UNIVERSITY OF STRATHCLYDE

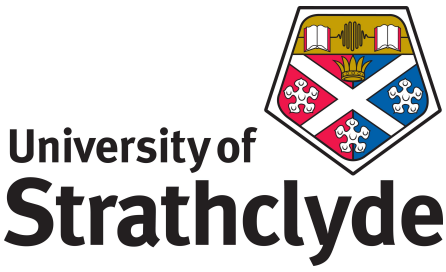
Department of Physics

Institute of Photonics

Smart Illumination Photometric Stereo based 3D imaging systems using LED technology

by

Emma Le Francois



A thesis submitted for the degree of Doctor of philosophy

September 2022

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Signed: Date:

Frontispiece

Picture showing the 'Top-down Photometric Stereo imaging' experimental setup with the 4 modulated LEDs illuminating the scene and the camera for the frame acquisition.

Abstract

The high-modulation rate of light-emitting diodes combined with a new modulation scheme called Manchester Encoded Binary Frequency Division Multiple Access enables the self-synchronisation of a set of white light-emitting diodes between themselves and a camera. Based on this smart illumination system, a new synchronisation-free photometric stereo imaging system is presented to achieve high-resolution 3D shape reconstruction in order to enhance indoor video surveillance systems.

This thesis first demonstrates the experimental proof-of-concept of the top-down illumination photometric stereo imaging system using off-the-shelf equipment such as commercially available white light-emitting diodes and a smartphone. A depth resolution of 3 mm for an object imaged at a distance of 42 cm and dimensions of 48 mm is reported. Dynamic imaging application of the experimental setup achieved a full 3D reconstruction of an ellipsoid in motion at a video rate of 25 fps with an error ranging between 4 mm and 11 mm at a similar distance.

A hybrid imaging system based on time-of-flight and photometric stereo imaging is also reported in this work. The former can achieve depth accuracy in discontinuous scenes and the latter can reconstruct surfaces of objects with fine depth details and high spatial resolution. The experimental proof-of-principal shows a root mean square error ranging between 4% and 5% for an object auto-selected from a scene imaged at a distance of 50 cm to 70 cm.

Finally, the generation of a bespoke synthetic dataset for top-down illumination photometric stereo imaging is achieved. This dataset aims to constitute the building blocks of a potential convolutional neural network which would decrease the computational time of the global image processing in order to reach real-time imaging applications.

Acknowledgements

First, I would like to thank my supervisor Prof. Martin Dawson, for the opportunity to undertake this PhD research and the support provided throughout. I would also like to thank Dr. Michael Strain for guiding my research, supporting my work and for answering my questions in my moments of self-doubt. Many thanks to Dr. Johannes Herrnsdorf for his ideas and work that helped me got started on the project and his help throughout the PhD. I would like to thank Dr. Jonathan McKendry for helping me on some electronic part of my work. Thank you to Dr. Alexander Griffiths for his help on the time-of-flight experimental setup.

From Fraunhofer UK, I would like to thank Dr. Adam Polak for his ideas in making the work more applicable and his support.

From Aralia Robotics in Bristol, I would like to thank Dr. Laurence Broadbent and Dr. Glynn Wright for their inputs on the photometric imaging work.

From the University of Edinburgh, I would like to thank Prof. Robert K. Henderson for leading me the Quanticam camera for the time-of-flight work.

I am grateful to EPSRC, Fraunhofer UK and quantum enhanced imaging (QuantIC) grants for providing the funding for my PhD research.

On a personal side, I would like to thank my family for believing in me; my friends for the laughs, great time spent together and their support that helped go through this challenge. I also thank all my colleagues at the Institute of Photonics for making every-day enjoyable. A special thanks to Pedro Alves, George Chappell, Joshua Robertson and Gemma Quinn for cheering up my spirits with all the great board game nights that we had! Finally, a warm thank you to my partner Michael Hutchinson for his support and love, whom helped me a great deal in finishing the PhD.

Contents

List of Figures

List of Figures

List of Figures

List of Figures

List of Tables

Abbreviations

2D Two Dimensional

3D Three Dimensional

APD Avalanche Photodiode

BRDF Bidirectional Reflectance Distribution Function

CCD Charge Coupled Device

CMOS Complementary Metal Oxide Semiconductor

CNN Convolutional Neural Network

DCR Dark Count Rate

EBL Electron Blocking Layer

FDMA Frequency Division Multiple Access

FM Fast Marching

FOV Field-of-View

FPGA Field Programmable Gate Array

GaN Gallium Nitride

GBR Generalised Bas Relief

GS Gram-Schmidt

LED Light-Emitting Diode

LiFi Light Fidelity

MEB-FDMA Manchester Encoding Binary Frequency Division Multiple Access

MQW Multi-Quantum Well

NRMSE Normalised Root Mean Square Error

OOK On-Off keying

PDE Photon Detection Efficiency

PDP Photon Detection Probability

PMT Photomultiplier Tube

PS Photometric Stereo

PS-FCN Photometric Stereo Fully Convolutional Neural Network

RMSE Root Mean Square Error

SNR Signal to Noise Ratio

SPAD Single-Photon Avalanche Photodiode

TCSPC Time-Correlated Single-Photon Counting

TDC Time-to-Digital Converter

TDMA Time Division Multiple Access

ToF Time-of-Flight

VLP Visible Light Positioning

Symbols

I_{mg}	measured image
ρ	albedo (reflectivity of each pixel)
N	surface normal components
L	lighting vectors
x, y, z	coordinates of four LEDs
I	intensity of the incident LED light at the object location
I_0	LED emitter intensity
A	Albedo image
Z	Height function
ϵ	quadratic function (error function)
W	Eikonal equation solution
λ	constant of the Eikonal equation (set to 0.06)
f	square Euclidean distance function
$RMSE$	root mean square error
$NRMSE$	normalised root mean square error
n	number of data pairs
d_i	difference between measured values and reference values
$s_{i,j}$	transmitted signal matrix
$\otimes$	Kronecker product
T	sampling period
v_i	frequencies of the square wave

I_i	intensity received from emitter i
r_j	sum of contribution from all emitters
D	decoding matrix
P_{signal}	optical power of the LEDs in Watts
P_{noise}	optical power of the ceiling in Watts
SNR	signal to noise ratio
h	image height
w	image width
L_{depth}	loss function
$\beta 1$	Adam parameter 1
$\beta 2$	Adam parameter 2
α	learning rate

Symbols

Chapter 1

Introduction

In this thesis, a new light modulation scheme, called Manchester Encoding Binary Frequency Division Multiple Access (MEB-FDMA), is introduced and applied in a Three Dimensional (3D) imaging scenario where its synchronisation-free and flicker-free properties are highlighted. The MEB-FDMA scheme is an orthogonal modulation scheme which is self-clocking and means that no trigger is required to start the acquisition process with a camera. In addition, the modulation of the light sources is fast enough so that the human eye cannot see the light flickering, which is effectively like having the scene fully lit at all times. The MEB-FDMA synchronisation-free feature; which self-synchronise both the camera and the lighting sources, but also the lighting sources between themselves; offers an easily deployable 3D imaging system that can be retrofitted in public spaces to achieve 3D reconstruction of scenes in order to enhance current video surveillance systems. The thesis focuses on the demonstration of a high-resolution 3D imaging method that relies on the smart application of general lighting using commercially available light-emitting diodes and a smartphone. Throughout the thesis, two main 3D methods will be applied to achieve high-resolution 3D reconstruction of a scene, namely the Photometric Stereo (PS) imaging and the Time-of-Flight (ToF), the latter will assist the PS imaging approach in discontinuous scenes. In other words, the focus of this thesis is to demonstrate that a video surveillance camera system can be improved by using already existing general lighting sources in a smart way to retrieve enough information of a scene to reconstruct it in 3D without impacting the

comfort of the users. The idea is to keep the deployment of the proof-of-concept fast and easy to implement which leads us to pick a robust 3D imaging technique known as photometric stereo imaging.

This introductory chapter first covers a general background on 3D imaging, followed by a description of the state-of-the-art 3D imaging techniques with a focus on photometric stereo, time-of-flight and sensor fusion. The goal of this chapter is to bring enough background information to the reader. The second part of the introduction will define camera properties and its limitation in different scenarios. This work will show that the imaging method can adapt to smartphone as well as more developed scientific cameras.

Following the introduction, Chapter 2 gives a quick introduction to the technical background required to understand what both a light-emitting diode and a single photon avalanche photodiode detector are and how they can respectively achieve a fast modulation and a fast detection to the single photon level.

Chapter 3 describes the mathematics of PS imaging along with a full explanation of the numerical method that integrates the surface normal vectors, called Fast Marching. In this chapter, the Two Dimensional (2D) Fast Marching code from Dr. Juan Cardelino has been extended to perform surface reconstruction using different properties and assumptions when compared to the original work, see more information in the chapter. The new Fast Marching algorithm that is presented in this thesis is robust to different integration borders which mean than any shape can be reconstructed. In addition, the algorithm takes about a minute to output the topography of the imaged object. This third chapter also mathematically explains the light modulation scheme called MEB-FDMA, with some direct examples on how the scheme is experimentally applied. The MEB-FDMA has been introduced by Dr Johannes Herrnsdorf in the research group and the work presented here is based on applying this already implemented modulation scheme to a 3D imaging application to highlight its main properties, which are the synchronisation-free aspect where the camera and the light sources, and the light sources between themselves, don't need to be synchronised together; and the light flicker being above visual flicker recognition.

In turn, Chapter 4 introduces the experimental work on synchronization-free top-down illumination photometric stereo imaging which demonstrates the proof-of-concept of the illumination modulation scheme applied with photometric stereo to reconstruct the topography of objects. This one of the main contribution in this thesis as this is the first time that a top-down scenario in PS is demonstrated. Current PS scenarios are limited by their deployability as the camera is always surrounded by the light sources which limit the application of the imaging system. While here, thanks to the MEB-FDMA scheme, a hand-held smartphone is used independently to the LEDs which are mounted to a ceiling, and a similar reconstruction resolution is reached.

Chapter 5 discusses whether the top-down illumination and modulation scheme approach can be adapted to reconstruct the shape of an object in motion, with an end-goal of achieving real-time imaging. In this chapter, the PS imaging setup is adapted into a dynamic setup by adding a stepper motor to rotate an object on a single axis and show that it is possible to reconstruct the shape of the object without synchronising the stepper motor and the camera.

In Chapter 6, the focus is on dealing with scene discontinuities and therefore presents a solution where time-of-flight imaging is used as a masking tool to select the object to reconstruct, and PS imaging is applied to this local area to reconstruct the object shape in high-resolution. This chapter also introduces the use of another sensor: a single photon avalanche photodiode detector that has timing capabilities to achieve time-of-flight measurement. This chapter is a shared contribution between myself and Alexander Griffiths as he set up the ToF work and I implemented the calibration algorithm to align two cameras and achieve high-resolution of auto-selected objects.

Chapter 7 presents the background of a deep-learning computational method where the aim is to improve the computational reconstruction time of the 3D scene in order to achieve real-time imaging. A contribution in this chapter is the generation of a new dataset, which contain PS images of about 15,600 different shapes and their corresponding topography. This dataset can be used in the future to train a deep-learning model where the topography of an object would be retrieved without having to determine its surface normal beforehand.

Finally, Chapter 8 summarises the findings of this thesis, and presents some avenues for further research and development. The code generated in this work for the top-down illumination imaging, the dynamic imaging, the hybrid imaging and the deep learning are available on the strathcloud code repository under this reference [19].

1.1 Three-dimensional imaging

Three-dimensional (3D) imaging is a broad area of research and can range from space-related applications to cell-level reconstruction of living organisms in the microscopic field. It is therefore important to define the research focus which will help direct the work carried out in this thesis.

Computer vision is an artificial intelligence research area that enables computers and systems to derive meaningful information, mostly based on reconstructed 3D imaging, obtained with a digital images, videos or other data, to then take actions based on the information. The main goal of computer vision is to help the computer "see" its surrounding environment like a human would using stereoscopic vision and a brain. The issue is that a computer does not process the two-dimensional images the same way as a human brain and for this reason researchers are looking to find algorithms that will give the tools to the computer to process in 3D the information to better sense the environment.

In computer vision [20,21], 3D imaging has been intensively focused on stereo vision systems that rely on a pair of images to compute the 3D scene just as in human vision. Although this technique is robust and well-defined, more research shows that other 3D imaging system can achieve faster and higher-resolution thanks to the development of high-resolution sensors and smart modulation of illumination systems. Based on the illumination technique, 3D shape acquisition can be divided into two categories: passive and active systems [22]. A passive technique is based on the optical appearance of an object under non-tightly focussed illumination coming from distant light sources [1], while an active technique examines the interaction (presence, absence, deformation) of the light and the surface with an energy signal e.g. from a laser [1]. Fig. 1.1 shows the taxonomy of the active and passive 3D shape acquisition techniques. A

more detailed state-of-the-art description of both approaches will be developed in the subsection below.

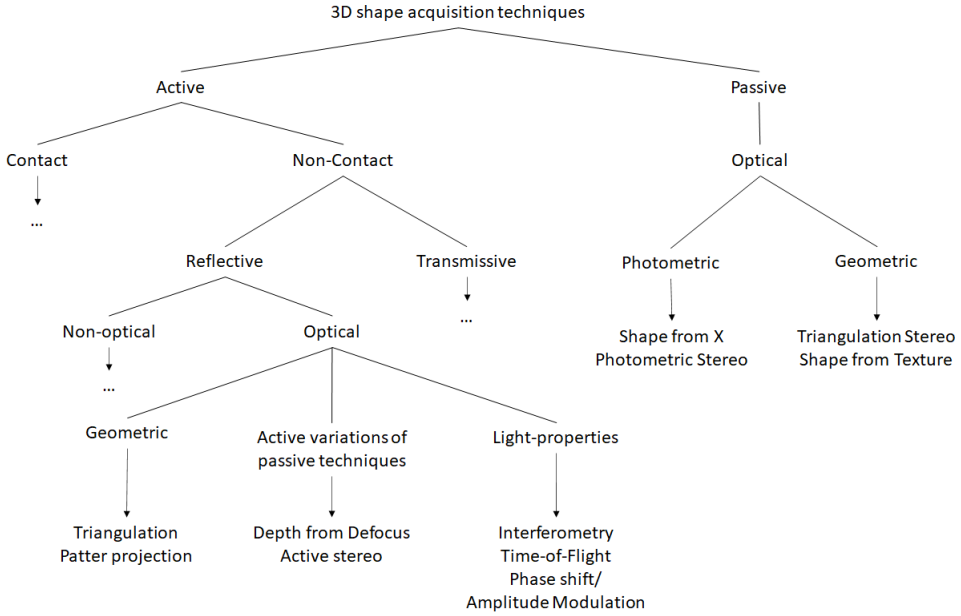


Figure 1.1: Taxonomy of active and passive 3D shape acquisition techniques [1].

In addition to 3D imaging being mainly based on the shape reconstruction technique, the choice of the sensor and the illumination system is equally important. Depth data-reconstruction is a very complex problem because of the object and its environment but also because of the sensor that is used for the data acquisition [1]. For example, the sensor may capture different data-type of 2D images (chromatic, monochrome, polarized light, etc.) or capture data under specific response curves and may be single/multiple acquisitions and may be fixed or moving. On top of that, the illumination may consist of one or several known/unknown light sources which can be distant or in the near-field related to the object, and the source can be a laser or a simple light emitting diode also used for general lighting [1]. All these variations gives an idea of the difficulties of depth reconstruction but also of the opportunities and the range of assumptions that can be derived regarding the approach chosen. Similarly, depth reconstruction encompasses different stages to obtain the final topography of the object, namely: the storage, the analysis, the transmission and the visualization of the 3D

shape [22].

An effective 3D computer vision approach opens wide new possibilities in different domains such as object recognition, 3D remote sensing, industrial quality inspection, scene surveillance and monitoring (which allows automatic recognition of unexpected behaviour) and robot navigation to name but a few. Much more information can be derived from a 3D scene than from a 2D projection obtained with a standard camera which helps define powerful algorithms to make decisions based on the depth reconstruction. As discussed in the previous paragraph, a 3D imaging or shape acquisition system can be active or passive depending on the use of the illumination source. Active 3D imaging includes techniques such as ToF, Lidar and structured-illumination, while stereovision can be applied to both and photometric stereo imaging is a passive technique. All of these methods are widely used in the current research as each method has its own particular merits and limitations regarding speed of operations, hardware required, depth resolution, etc. For example, ToF, which is an active method due to the interaction of the laser beam with the scene, is recommended for outdoors' long-range target recognition as powerful sources are needed to receive enough photons back on the sensor [23], [24], while a structured-illumination technique would be more adequate for indoors application for its faster acquisition time compared to a raster scan technique [25], [26]. Other studies show that methods can be merged together to achieve fast long-distance measurement. For example, a photon-counting computational LiDAR system can utilize a short pulsed structured illumination combined with a fast response photomultiplier tube for reconstructing 3D scene at up to 3 fps [27]. Nonetheless, a main issue with active 3D imaging is its difficulty in reconstructing objects in motion [1]. Passive methods rely on the optical appearance of an object under non-focused illumination [1] which gives the advantage of not interfering with other sensor devices.

1.1.1 Stereovision

Stereoscopic imaging has been widely used for years for robot navigation [28], [29] and is a passive triangulation method that can deduce the depth information of a scene

from multiple static 2D images. In the classical stereoscopic vision technique, called stereo vision, two cameras acquire stereo-pair images, each obtained from a different viewpoint in space, to then decode the depth information which is implicitly captured in the stereo-pair. The advantage of stereovision is that it does not require any dedicated light sources.

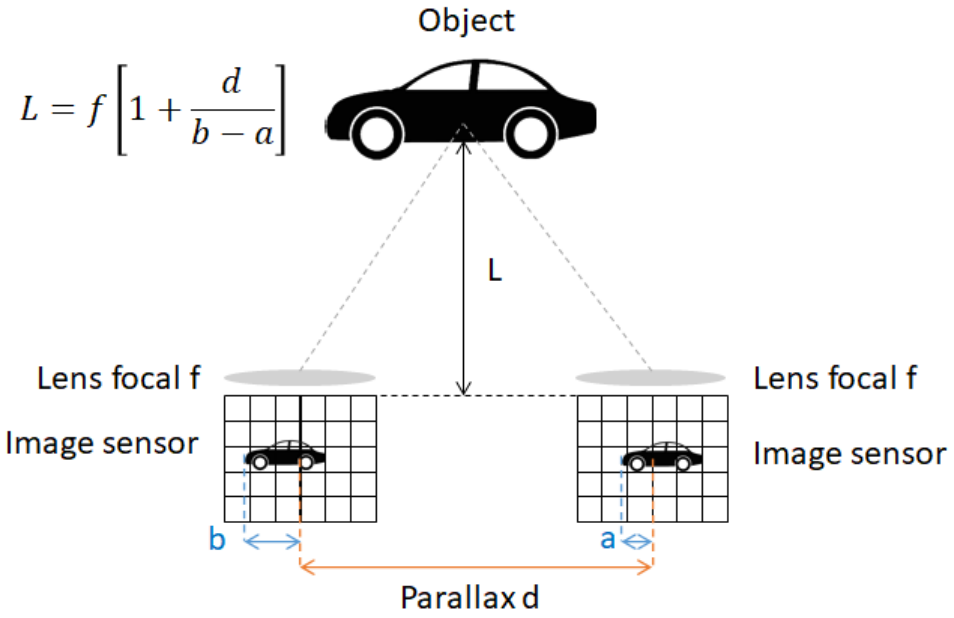


Figure 1.2: Binocular stereovision schematic. Two cameras are displaced from each other with a parallax d. The disparity b-a and the focal length allows determination of the distance to the object.

In practise, two cameras are displaced from each other, see Fig. 1.2. By knowing the camera focal lengths and the geometry, the depth of objects in an imaged scene can be estimated in a canonical stereoscopic vision system [30]. From the captured images, the relative displacement of the contents of one of the images is measured with respect to the other image, known as the parallaxes [21]. The range image of the scene is then obtained in two different steps. First, the correspondence process consist of a searching and matching technique to find pairs of matched points in the acquired two images. For these points to match, their projections in the scene must correspond to the same 3D point. Different local and global matching algorithms can

be used to produce the disparity map such as region-based, features-based or phase-based stereo matching algorithms [31]. In a few words, the matching methods will have different starting points. For the region-based algorithm, a point in the reference image is selected and then a support window in the neighbourhood of the selected point is obtained. The goal is then to find a sub-window that is similar to the support window in the image to be matched according to certain similarity judgement criteria. The pixel point corresponding to the sub window is the corresponding matching point [32]. The output of this correspondence process is a disparity map, more or less dense depending on the matching algorithm used, which represents the difference of the matched points on the horizontal coordinates [30]. The second step is the reconstruction process which is based on the disparity map and the stereo geometry of the scene.

Both active and passive illumination can be used in stereovision. Active illumination for stereovision means that some patterns are projected onto the scene to facilitate the finding of the parallaxes between the stereo-pairs images, which can be an ambiguous task. Projected patterns often comprise grids or stripes and can sometimes be colour-coded [21]. Stereovision mostly relies on the accuracy of the matching algorithm which means that the main challenge of stereo computation lies in the design of the corresponding searching algorithm between two measured images. This method requires a certain amount of time to create the disparity map that contains the matching pixels from both cameras' perspective [30].

Although new low cost high-resolution synchronised cameras can be used to improve the acquisition time, the camera's calibration and its synchronisation limit the deployability of this technique. In addition, stereovision is dependent on the scene textures which makes it difficult to provide reliable information under varying lighting conditions [30], hence it is less robust to different illumination scenarios occurring in real-life application. Another drawback of stereovision is the shadowing effect which can be minimised by using multi-view triangulation systems at the price of an enormous increase of data processing as well as increasing the number of cameras.

1.1.2 Structure-from-motion

Structure-from-motion is in a way closely-related to stereovision as both methods exploit image correspondences across multiple viewpoints of the same scene to obtain 3D information [33]. However, structure-from-motion is not limited to two frames like stereovision and makes use of the relative motion between camera and scene by extracting the shape of a scene from the spatial and temporal changes occurring in an image sequence [34]. Structure-from-motion has been applied in a variety of fields such as robotics, augmented reality, architecture, archaeology and visual inspection [33]. Instead of a single stereo pair, this technique requires multiple, overlapping images to achieve feature extraction and 3D reconstruction algorithm [35] using robust visual correspondences. The differences between consecutive frames are, on average, much smaller than those of typical stereo pairs, because image sequences are sampled at high rates. By using triangulation methods of the location of multiple 3D points, a sparse representation of the scene is formed and the intrinsic/extrinsic parameters of the camera from every viewpoint is calculated [33]. In general, structure-from-motion techniques are divided into two main stages: correspondence searching and reconstruction as shown as an example in Fig 1.3.

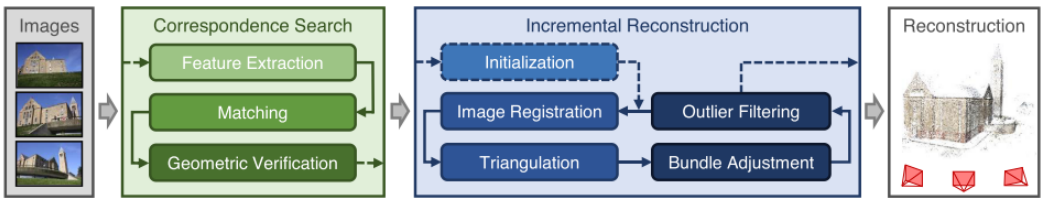


Figure 1.3: Structure from motion pipeline reconstruction [2].

Regarding correspondence search, the fact that motion sequences provide many closely sampled frames for analysis is an advantage. Firstly, tracking techniques, which exploit the past history of the motion to predict disparities in the next frame can be used. Secondly, the correspondence problem can also be cast as the problem of estimating the apparent motion of the image brightness pattern such as the optical flow. Two kinds of methods are commonly used to compute the correspondence. Differential methods use estimates of time derivatives and require therefore image sequences

sampled closely. This method is computed at each image pixel and leads to dense measurements. Matching methods use Kalman filtering to match and track efficiently sparse image features over time. This method is computed only at a subset of image points and produces sparse measurements [34].

Unlike correspondence, reconstruction is more difficult in structure-from-motion than in stereo. Frame-by-frame recovery of motion and structure turns out to be more sensitive to noise. The reason is that the baseline between consecutive frames is very small. For reconstruction, the motion field of the image sequence can be used. The motion field is the projection of the 3D velocity field on the image plane. One way to acquire the 3D data is to determine the direction of translation through approximate motion parallax. Afterwards, a least-squares approximation of the rotational component of the optical flow is determined and used in the motion field equations to compute depth [34].

1.1.3 Photometric Stereo imaging

PS imaging is a passive illumination imaging technique that has the advantage of being more accurate than stereovision for high frequency depth data and that can deal with objects in motion and in un-textured areas [1]. It has high-resolution [36] and fairly fast computational time, requiring only one camera to be calibrated, making it one of the most common 3D imaging techniques [37] for indoors scenarios. According to [38], PS imaging can have a higher resolution and can recover more highly detailed surface geometry than structured illumination or state of the art commercial laser scanners for common operating conditions.

First introduced by Woodham in 1980 [39], PS imaging relies on having one fixed camera perspective and different illumination directions to image an object in 3D. This technique determines the surface normal vectors and surface albedo at each pixel of the captured images assuming a perfectly diffuse (Lambertian) surface of the imaged object [39]. Surface normal components can then be integrated to recover the 3D shape. Conventional calibrated PS methods assume that the light directions/intensities are known or unknown but then identical across the acquired images [40]. The calibration of

the lighting directions can be demanding and requires accurate calibration methods [37, 41]. Knowing the position of the light is important to resolve the Generalised Bas Relief (GBR) [42]. If the light source directions are unknown then the structure of the object can only be determined up to a GBR transformation where shadows do not provide further information [42]. In addition, the position of the lighting source in relation to the scene will also have an impact on the assumptions made in PS imaging. The light source can be located far enough from the scene to consider a far-field scenario where the light will be diffused with a straightforward calibration. However, if the light source is closer to the scene to be reconstructed then a near-field scenario needs to be considered and a more accurate calibration will be needed [43, 44]. To avoid this time-consuming task, researchers are investigating semi-calibrated [40] or uncalibrated PS methods [45–47]. Moreover, PS imaging research is also broadening to non-Lambertian surfaces where the study of the Bidirectional Reflectance Distribution Function (BRDF), which gives the general description of the reflectance property of a surface, allows the consideration of real-world objects [48–52]. The PS imaging research area is broad and the work in this thesis will focus on finding a way to apply PS imaging easily and quickly, with a simple calibration method of the light directions.

In the literature, some research in PS imaging has been focused on the deployment of the method for the 3D reconstruction of bodies, faces or objects at low-cost. In 2009, a full human body dynamic shape capture was achieved by Vlasic et al. [3] using multi-view PS imaging, see Fig. 1.4 a). By using 8 to 9 different cameras and a dome of lights, they retrieved detailed geometry information of the human body in motion with an accuracy of few millimetres and a temporal resolution of 60 Hz. Although Vlasic's imaging method reproduces impressive detailed geometry in motion, the set-up design is not easily deployable in public areas. For the face reconstruction, Schindler's technique of using computer screen patterns to illuminate a human face for real-time reconstruction with PS imaging [53] shows the feasibility of using non-expensive and widely available devices with PS imaging. There is some limitation to this method though as it needs a dark room to work properly. More recently, Salvador-Balaguer in [5] demonstrated a non calibrated PS imaging method reaching a few millimetres ac-

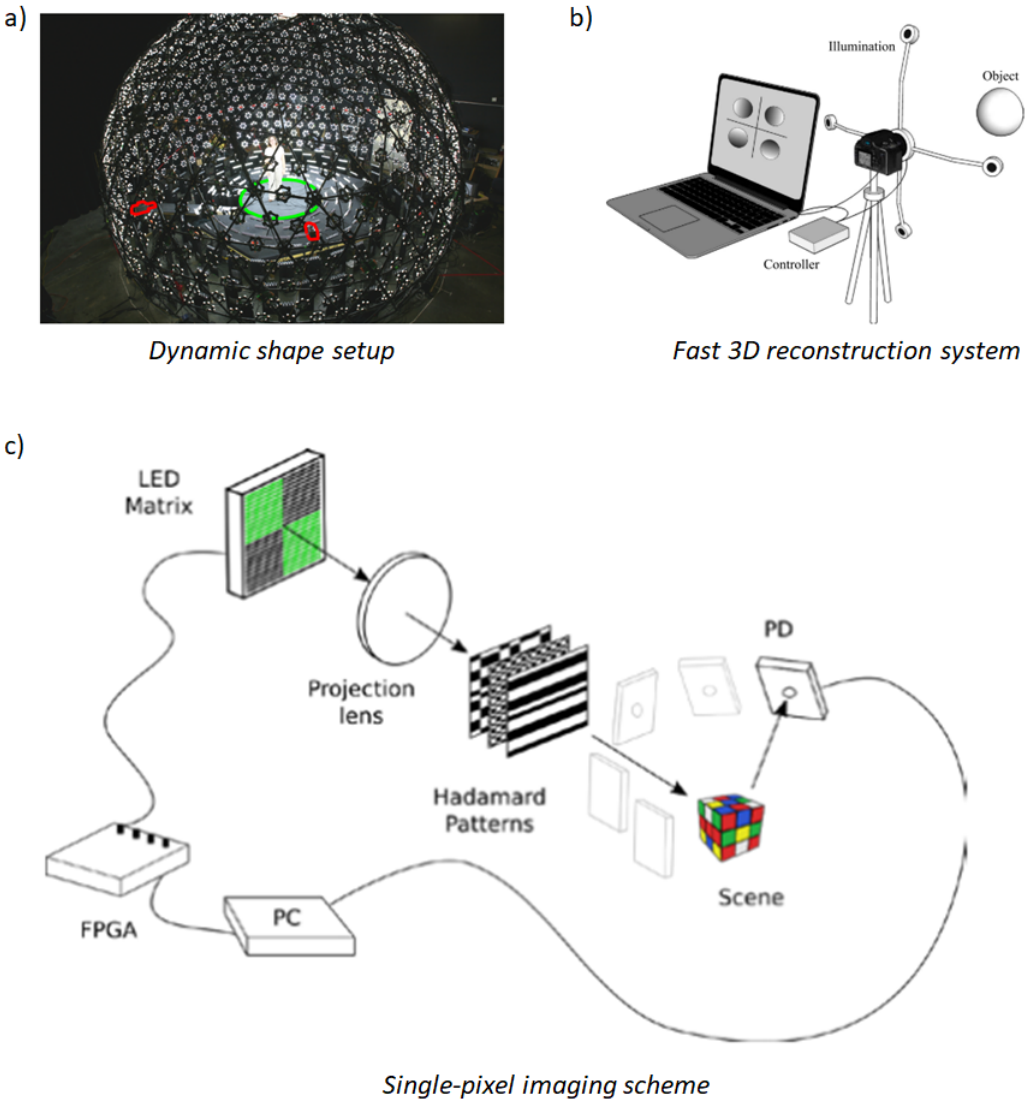


Figure 1.4: Photometric stereo imaging experimental setup from the literature. a) Dynamic Shape capture using multi-view Photometric Stereo [3]: 1200 individually controllable light sources, 8 cameras, with an additional ninth camera looking down from the top of the dome onto the performance area. b) Fast 3D reconstruction system with a low-cost camera accessory [4]: four LEDs, fixed around the camera lens and connected to a electronic controller board, both camera and controller board are controlled by a program on the laptop. c) Single-pixel imaging scheme based on a colour LED display [5]: sequence of Hadamard patterns displayed on the LED panel, light collected by a Photodiode and digitized by a DAQ system controlled with a computer, the photodiode is shifted to different positions to achieve 3D imaging.

curacy at a 2D resolution of 32x32 pixels. In this experiment, displayed in Fig. 1.4 c), a low-cost colour Light-Emitting Diode (LED) array controlled by a Field Programmable Gate Array (FPGA) projects patterns onto the object and the light reflected is collected by a single photodiode. To achieve 3D reconstruction, the single photodiode was shifted to different positions around the object. The method is similar to PS imaging but the illumination is fixed and the camera moved to obtain different shades of the object. The main issue here would be the "ready-to-use" aspect of the technique as a moving camera would not fit a video surveillance configuration. Ideally, to make PS imaging user friendly, a fixed camera perspective and some fixed illumination sources are needed. Zhang in [4] demonstrated a 3D reconstruction system using a low-cost camera accessory. They implemented a commercial camera that is placed in front of the object with at least four white LEDs surrounding it in a top/bottom/left/right configuration, see Fig. 1.4 b), adapted to X shape fashion in [38] and [54]. A fast 3D reconstruction has been reported with white LEDs surrounding a camera, where LEDs were sequentially lit by a USB programmable board [4]. These 3D reconstructions showed a standard deviation error ranging from 2.65 mm to 15.60 mm for objects of size 50 mm and 160 mm, respectively.

A main issue in PS imaging is the discontinuity, *i.e* the rapid change in intensity between an object and the background which results in a gradient on the depth axis close to zero hence a high value artefacts in the surface reconstruction problem. These high artefacts create discontinuity problem in the surface reconstruction which makes the surface normal of the object difficult to integrate. The state-of-the-art work cited above reports 3D reconstruction of a single object. The PS method does not allow the reconstruction of multiple objects within a scene because of high values artefacts arising in the gradient system of equations. To tackle discontinuity issues in PS imaging, methods such ToF are considered. Specular reflections and shadows are other limitations that can severely impact the robustness of the PS imaging method [55].

1.1.4 Time-of-Flight

Time-of-flight (ToF) is another 3D imaging technique, based on the same method as light detection and ranging (LiDAR), which is most suited for long range target recognition. The first ToF system, based on low-resolution sensors, could not achieve high enough resolution to be widely spread. Now with the recent advances in Single-Photon Avalanche Photodiode (SPAD) [56] and the implementation of laser-based ToF [23,57] or LiDAR [27,58] high-resolution at remote distances is now achieved [24]. This technology advance has made it possible to use of ToF in different applications such as machine vision, navigation of autonomous vehicles, and spacecraft, and atmospheric remote sensing [59,60].

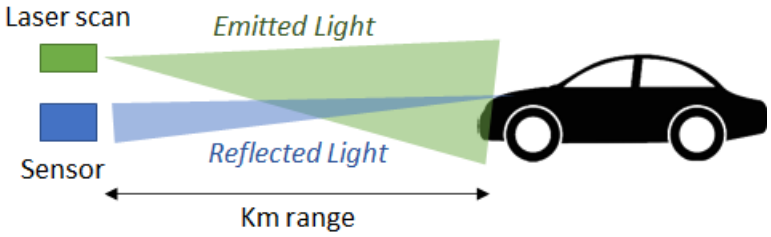


Figure 1.5: Schematic of Time-of-Flight, ranging and imaging.

To obtain a high-resolution depth map, a short-pulsed laser illuminates a scene and a sensor measures both the time and the intensity of the reflected pulse, where timing electronics time-correlate the reflected light-delay time and the intensity with the outgoing pulse [57], see Fig. 1.5. The literature reports laser-based ToF systems that can reach a 3 mm accuracy at a range of 5 m [61]. Much longer distances have been reached, such as 10 km with mm range resolution [23]. However the range is not the highest difficulty in ToF imaging as the amount of light detected by a sensor is the real challenge. Time correlated single photon counting techniques make it possible to reconstruct a signal with average signal returns of less than one photon per pixel [23]. Recent advances in SPAD detectors show that high-sensitivity, low-noise devices can be combined with dense logic [56]. Each pixel of a SPAD array can extract a distance measurement from a light pulse emitted on a scene and reflected back on the sensor [56,

62]. ToF imaging deals well with long range and discontinuous scenes but is limited by the resolution of current single photon camera systems, or the acquisition time of scanning systems [63, 64]. Very narrow pulses must be sent to retrieve mm depth map resolution, as demonstrated by the simple equation $d = c \cdot t$ where d is the distance, c the speed of light and t the time-delay. For example, to achieve a 1 mm depth resolution, a pulse of 3.3 ps must be sent by a laser. A drawback of ToF for extremely accurate ranging is the need for ultra-short pulses which requires significant pieces of hardware such as solid-state lasers which are bulky and expensive.

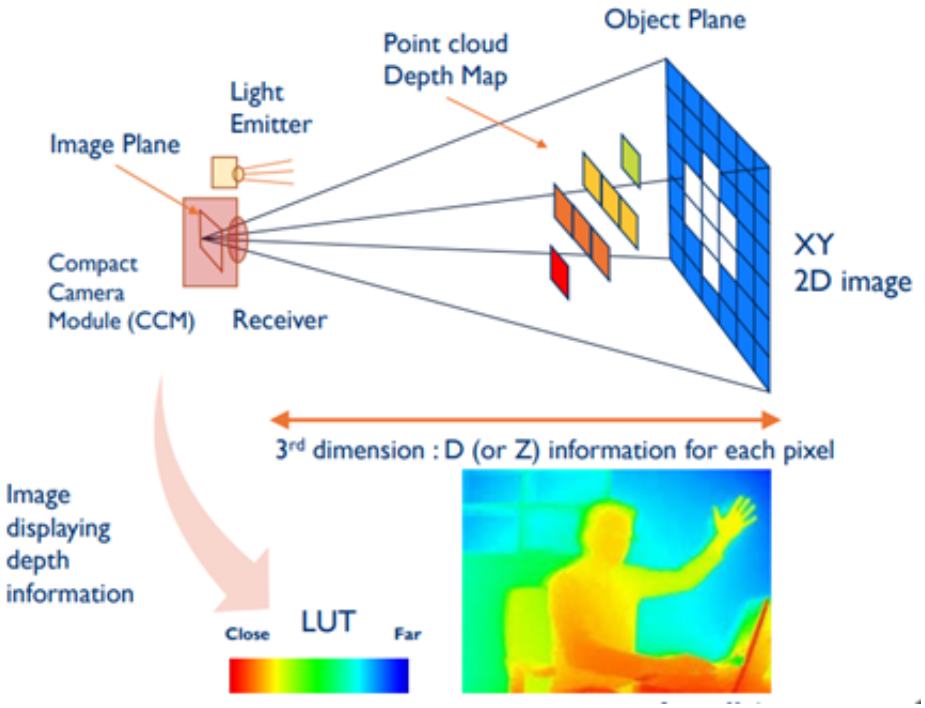


Figure 1.6: Schematic representing the principle of a depth camera [6].

LED-based ToF can also be used to measure the depth map of a scene but at a lower resolution as ultra-short pulses are not yet achievable with LEDs despite their high-modulation rate [62]. This technique opens great possibilities for implementing cheap and small ToF devices that would be resistant to discontinuities despite a lower resolution which would be in the cm range. It is also important to highlight that more and more depth cameras are now commercialised and ready to use such as the Kinect [65] or the blaze camera from Basler [66]. As shown in Fig. 1.6, the physics

behind a depth camera remain the same as a laser-based camera. To be resilient against background light, an infra-red light pulse is sent onto the scene to image it and each pixel of the depth camera will measure the time elapsed between the outgoing pulse and the received pulse. Each pixel with time information is then converted to an image that contains the depth information, such as a colour coded point cloud. In addition, it is now possible to come across ToF sensors being implemented onto Ipad and Iphone which are widely available to the public. As shown in Fig. 1.7, the new Ipads and Iphones use a ToF camera as a complementary sensor to improve the picture quality in low-light scenarios or measure the 3D map of a scene for architecture application. Recently, Luetzenburg's group evaluated the iPhone 12 in terms of accuracy for geoscientific applications and reported an absolute accuracy between 1 cm and 10 cm respectively for small object of 10 cm and for a 3D model of a coastal cliff with the dimensions of up to 130 × 15 × 10 m [67].

Figure 1.7: Picture of the Iphone 12 showing the integrated LiDAR sensor (left) [7], a 3D room scan from Occipital's Canvas application, enabled by depth-sensing LiDAR on the IPad Pro (right) [8].

Laser-based TOF is preferentially used for space applications where solid-state lasers can be retrofitted onto satellites to then generate high-power-short pulses to map the topography of the Earth [60], or for autonomous vehicles with different pulse width [68]. The reader can refer to the literature for more information and details on high-resolution ToF systems [23,24], LiDAR [58,68,69] and the new challenges arising with depth cameras [63,64,70–72]. A more suitable approach to this thesis work is LED-based ToF as it would help overcome the discontinuity issues of the PS method while keeping the technology at a low-price and remaining easy to implement without requiring high-powers.

However, because of the LED properties and the achievable pulse width (ns), the depth map resolution will not achieve the same accuracy as the PS imaging method. Research has shown that depth cameras can be merged with other 3D imaging application or colour cameras to enhance the accuracy of the depth map. This technique is called sensor fusion.

1.1.5 Sensor Fusion

Sensor fusion has been investigated in the past few years to mainly improve a ToF camera's low-resolution depth by using high-resolution intensity images [73]. Sensor fusion can also be used to improve 3D mapping procedures with colour-depth (RGB-D) cameras [74, 75] or perform segmentation and tracking [76]. Because of the impact of random noise [63], fine depth details are lost when using ToF cameras. However, PS imaging is robust to noise and can provide finer depth details than ToF [9], but PS does not provide absolute distances. By fusing devices together, the technique can take advantage of the spatial redundancy, while another approach would be to use image super-resolution techniques to utilize temporal redundancy [77]. Most of current sensor fusion techniques rely on improving the ToF sensor depth map by measuring both a range map and a surface normal map and then merge these to improve the resolution [9, 38, 77].

For example, the work of Sun Kwon Kim [78] merges the PS and the ToF imaging into one single device, see Fig. 1.8. Instead of having two sensors, one colour camera for the PS imaging and a depth camera for the ToF, the setup has been rearranged to use the ToF camera as the main camera for both methods. By controlling the turn-off patterns of the infra-red LEDs of the camera, both depth and normal maps are obtained simultaneously and then combined. The work reports an absolute difference of the ground truth and the measured scene that has decreased from 4.57 mm to 3.77 mm [78].

In this thesis, Chapter 5 will show that to improve the PS imaging method, the range map obtained with the ToF will be used as a mask to deal with discontinuities. One advantage of this is that the fusion scheme technique will not impact the PS

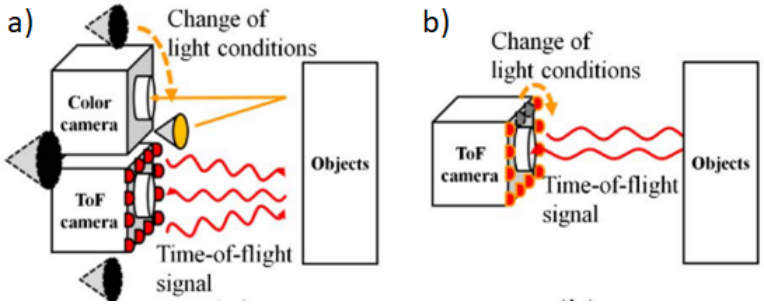


Figure 1.8: Sensor fusion schematic example [9]. a) Previous work where two cameras are used to acquire a surface normal map and a depth map. b) Current work where both the photometric stereo and the time of flight imaging are merged together using one depth sensor and infra-red LEDs.

imaging procedure and that only a calibration of both methods will be required to merge both maps. By employing a dual imaging system incorporating both ToF and PS, the complementary properties of both systems can be used to image complex 3D fields with high-resolution and complex discontinuities between objects.

1.2 CCD and CMOS cameras

Modern camera technology relies on both Charge Coupled Device (CCD) and Complementary Metal Oxide Semiconductor (CMOS) image sensors, which differ in terms of the image quality, the noise level and the global cost. CMOS cameras are much less expensive to manufacture but does not provide as high quality image as the CCD, although CMOS sensors are rapidly improving in performance.

Both technologies convert light into electrons by capturing light photons with wells called photosites. When an image is being taken, the photosites are uncovered to collect photons and store them as an electrical signal. An image is formed by accumulated the charge of each photosite. The main difference between a CCD and a CMOS camera is on the way these charges are read out in each photosite. In a CCD device, the charge is transported across the chip and read at one corner of the array where an analog-to-digital converter turns each photosite's charge into a digital value. While, in a CMOS device, there are several transistors at each photosite that amplify and move the charge using traditional wires. As each photosite can be read individually, it gives

more flexibility to apply the image sensor for different applications.

However, a special manufacturing process gives CCD devices the ability to transport charges across the chip without distortion, leading to high-quality, highly sensitive sensors. On the other hand, CMOS chips are manufactured with conventional and cheaper processes. The manufacture process directly impacts the image quality as CMOS sensors are usually more susceptible to noise while CCD sensors generate high-quality, low-noise images. This is due to the individual transistors mounted under each photosite that are sensitive to the photons that can hit the transistors instead of the photosite, which impact the light sensitivity which is lower in a CMOS sensor.

Based on these properties the camera choice for scientific applications is important as a few parameters have to be considered such as the image resolution, the linearity of the sensor, the image quality, the noise level and also the power consumption which is not negligible. This thesis demonstrates a proof-of-concept for an easily deployable, low-cost, 3D image reconstruction process where power consumption should also be considered. CCD sensors consume as much as 100 times more power than an equivalent CMOS sensor and a lower power consumption might be preferable.

1.2.1 Linearity response in cameras

An important parameter to consider in imaging sensors is the linearity response. For a camera to be linear in term of its response, the signal output should ideally be linearly proportional to the amount of light incident on the sensor after the digitization process. In most commercial cameras or smartphones, gamma correction is added to code luminance into a perceptually uniform domain which will introduce non-linearity to the image sensor [79]. The issue with smartphone, for example, is that the gamma correction is applied within the sensor and is like a black box where it is not possible to acquire the raw data. Linearity can become an important parameter in image analysis for shading correction, linear transforms [79] or medical applications [80].

A study on the linearity analysis of a CMOS sensor [79] shows that high-performance CCD image sensors can achieve excellent linearity with a non-linearity as low as a few tenths of a percent. While a CMOS camera reports a non-linearity of several

percent [81]. Research are being carried out to reduce this non-linearity issues in CMOS sensors [81].

A non-linearity response can also be caused by a saturation of the sensors with incident light. Although CCD sensors respond in a linear manner over a wide dynamic range, when full well conditions are reached under high average illumination intensity, a non-linear response is usually observed [82]. In this work, the exposure time of the camera, or the light source power, will always be adjusted so that saturation of the image sensors is avoided. However, as a smartphone will be use, it will not be possible to control the exposure time nor the linear response of the image as this will be integrated in the slow-motion mode of the smartphone. Nonetheless, results will show that the imaging system is robust to potential non-linearity response of the smartphone sensor and it will not impact the topography reconstruction of the object, which shows how adaptable and robust the 3D imaging system can be. No further investigation on linearity will be carried on in this work so the reader is invited to refer to the literature for further information on potential issues caused by non-linear image sensors [83].

1.2.2 Rolling shutter and global shutter

Another parameter to consider with an imaging sensor is whether an image is acquired in a rolling shutter operation (row by row) or a global shutter when all the pixels capture light at the same exposure time, see Fig. 1.9. In a rolling shutter operation, an object in motion can move faster than the exposure time of each row and result in an image that is stretched in space and which does not represent the reality. Commonly, modern CMOS cameras adopt the rolling shutter operation for its faster acquisition speed despite having to deal with some artefacts. For example, each row takes a certain amount of time to read out the pixels, let's say 10 µs, which is known as the line time. Now, as described in Fig. 1.9, if a camera has 2048 rows then the first row will be read at time 0 and the last row at time 2.048 ms and this is known as the frame time. It is clear that this small delay will impact the imaging of very fast samples [10].

On the other hand, scientific CCD cameras adopt a global shutter approach where

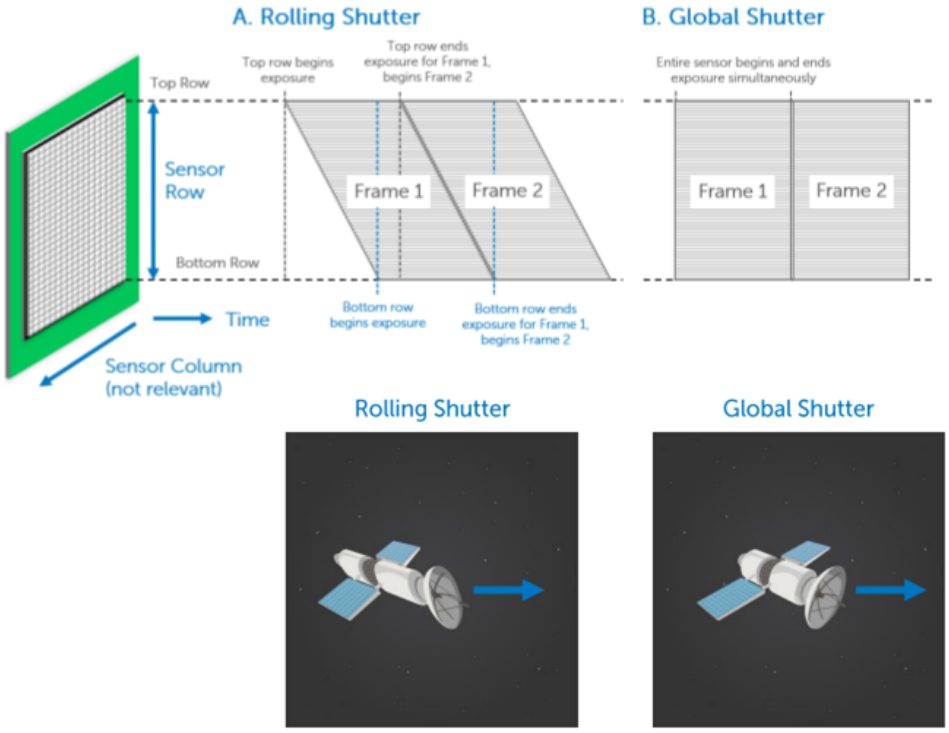


Figure 1.9: Acquisition timing diagram for rolling shutter and global shutter operation. An object moving quickly from left to right during the roll from top to bottom of a rolling shutter camera can appear skewed as the acquisition of the different rows occurs at different time during the object's motion [10].

all sensor pixels are read out simultaneously. The advantage of a global shutter is that the image obtained is a snapshot of a single point in time and this can be an important property where synchronisation between the camera and the light source is required by the use of a hardware trigger [10]. However, the readout of the full sensor is slow due to the camera having only one analog-to-digital converter. In addition, the more pixels on the sensors the slower the total frame rate of the sensor will be and this issue can then have an impact on read noise or longer duty cycle for the camera [84].

A trade-off between sensor's frame, image's artefacts, noise level have to be found depending on the application which will then impact the sensor choice between CCD and CMOS. For a full-review on the technology behind a CCD and a CMOS camera, the reader is directed to the following references [84–86].

1.2.3 Cameras used in the thesis

In this work, the first camera used is a smartphone Galaxy S9 with a CMOS rolling shutter. As explained earlier, some gamma correction are applied in smartphone and it is not possible to get the raw imaging pre-digitization. However, Chapter 4 will show that the non-linearity process applied to the image does not impact the determination of the surface normal and therefore the topography reconstruction of the object. In addition, the rolling shutter operation for now is not an issue as the scene is static. By using an off-the-shelf camera like the one built in a smartphone, it demonstrate that the PS imaging can indeed be robust to any kind of camera as non-linearities are not impacting the reconstruction.

In addition, another camera is used in Chapter 5 for the reconstruction of object in motion, the Fastcam mini from Photron. The camera choice is made on the memory capacity and not directly on the sensor technology. The camera provides a memory capacity of up to 32 GB. This high-speed camera is based on a CMOS sensor that can achieve high light sensitivity from a small image sensor (10 µm pixel pitch). At full-image resolution (1280 x 1024 pixels), the frame rates up to 4,000fps [87].

In short, camera sensors have to be carefully selected regarding the imaging application, especially when it comes to live microscopy or medical imaging. However, the

work done in this thesis gives the freedom to use any kind of camera to the point that it can achieve acquisition rate of at least 960 fps. In this case, a commercial CMOS camera is more likely to achieve higher frame rate and can therefore be considered as the main camera to be used.

1.3 Conclusion

To conclude, the main goal of the thesis is to report a proof-of-concept of an easily deployable, off-the-shelf devices, 3D imaging system that can improve in the future surveillance camera system in public spaces. In this introductory chapter, the main contributions of the thesis have been presented along with a detailed description of each chapter. A state-of-the art on 3D imaging methods have allowed to put in perspective the latest research works and figure out the best 3D imaging approach for this thesis. The 3D imaging methods overview on stereovision, structure-from-motion, PS imaging, ToF and sensor fusion, shows that computer vision have been greatly improved over the last decade and that it is possible to merge some of them to reach higher performances. Moreover, an overview of the CCD and CMOS showed that a choice has to be carefully made when it comes to selecting a sensor for scientific application as linear response or the shutter operation can have an impact on the resulted image. In the next chapter, a technical background on the physics behind and SPAD array will be given to support the understanding of the following work.

Chapter 2

Technical background

In this chapter, a technical background on the physics of light-emitting diodes and SPAD array is given to support the understanding of the PS imaging setup and the ToF setup that will be developed in Chapters 4 and 6, respectively. The reader will have a better vision on why LED can easily reach high-modulation rate and how a SPAD array detection is sensitive down to a single photon.

2.1 Light-emitting Diodes

Discovered by accident by Henry Joseph Round in 1907, the first light-emitting diode (LED), made of Silicon Carbide, emitted a yellowish light [88]. Years later, in the 1990s, Isamu Akasaki, Hiroshi Amano and Shuji Nakamura developed the first efficient blue-emitting LEDs made of Gallium Nitride (GaN). In 2014, this ground breaking discovery was recognised by a Nobel prize in physics for producing efficient blue LEDs [89]. Blue-emitting LEDs made possible the efficient operation of white LEDs, which is a key element in the Nobel prize citation. The majority of modern LED light fittings consist of a blue GaN based LED with a phosphor coating to convert some of the blue light to longer yellow wavelengths. White LED can also be obtained by combining a set red, green and blue LEDs to create a multichip. Fig. 2.1 a), b) shows, respectively, a blue-emitting and a white LED.

Light-emitting diodes are now ubiquitous in the modern world and are broadly

Figure 2.1: a) Blue LED, b) White LED and c) smart-lighting using LED bulbs.

used in applications where efficient optical sources are required, such as car headlights, general lighting, traffic lights or even TV screens. This technology is also known as smart lighting, see Fig. 2.1 c), and is, for example, made available in homes to control the lights remotely. By taking advantage of their fast modulation bandwidth which can range from several MHz to GHz [90], low voltage operation and convenient interfacing with digital electronics [91, 92], LEDs are now of interest for additional functionality such as data communications [11, 62, 93, 94] or object tracking and location [95, 96]. A new LED-enabled communication technology called Light Fidelity (LiFi), which means light fidelity, is being deployed as an alternative to Wifi and achieves indoor optical communication [97].

In this work, only white LEDs and blue LEDs are going to be useful, the latter to be used as a pulsed source for the ToF imaging. As the main goal is to retrofit the 3D imaging system to buildings and public spaces, it is important to use light sources that are widely used in most general lighting applications, and that are cheap and efficient. LED technology matches all these requirements. A quick overview on LED physics, optical properties and modulation rate is given in this section. For more details on the growth of LED materials, the spontaneous recombination processes, and a detailed overview of LED modulation characteristics, please refer to [11, 98, 99].

2.1.1 Semiconductor LED Physics

The physical mechanism that occurs when semiconductor LEDs emit light is spontaneous recombination of electron-hole pairs injected under bias and the simultaneous emission of photons, also known as electroluminescence [100]. The stimulated emission process that appears in semiconductor lasers is fundamentally different to the sponta-

neous process happening in LEDs. LEDs are fabricated from semiconductors, which are a class of material that contains a conduction band and a valence band, defining the allowed states of, respectively, electrons and holes (negative charge carriers and positive charge carriers). In the conduction band, most of the states are empty with conduction being carried out by the minority electron charge carriers. The electron band structure has a forbidden region called the "band gap" between the conduction and valence bands, which cannot be occupied by charge carriers. The spontaneous emission occurs when electrons from the conduction band recombine with the holes in the valence band, releasing energy as photons [11,100]. Depending on the energy difference between the band gap, the wavelength of the emitted photon will be different due to energy conservation [100]. Therefore, the desired emission wavelength of an LED can be attained by choosing a semiconductor material with an appropriate band gap energy.

Semiconductors with appropriate doping regions can be used to create p-n junctions. To increase the number of available carriers, semiconductor materials are doped with donor (n-type) or acceptor (p-type) atoms. When there is an excess of carriers and holes in the p-n junction, they diffuse into the opposite material type and recombine. This recombination produces a depletion region, also known as the active region, between the n-type and p-type material, which in turn forms a diffusion voltage V_D and produces a potential barrier of energy eV_D. Fig. 2.2 shows the energy band structure of the p-n junction. Under forward bias, carriers are injected into the depletion region where they can recombine and emit photons.

The density of carriers in the active region defines the radiative recombination rate in an LED, hence the amount of photons emitted. To control the emission wavelength of an LED, careful engineering of the band gap through material and width choice is necessary. In a p-n junction, electrons and holes propagate through a material with a characteristic diffusion length, which can be limited by the band gap global structure, hence limiting the carrier concentration. To increase the carrier concentration, the carriers need to be confined to a smaller region. To do so, large band gap materials are grown on either side of a narrow band gap material which forms potential

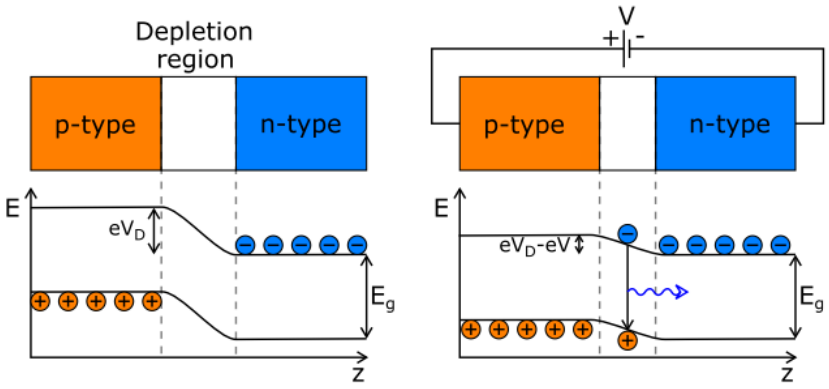


Figure 2.2: Energy levels in a p-n junction in equilibrium (left) and under forward bias (right) [11].

barriers that confine the carriers. If the active region width is comparable to the de Broglie wavelength of the electrons ($\approx 10\ nm$), quantum confinement effects become important [11, 100]. Such a so-called quantum well structure increases the energy of the emitted photon relative to the intrinsic material band gap energy. Indeed, by discretising the electron energy levels, the quantum well will raise the lowest energy levels just as the infinite potential quantum well case [101].

The goal in a p-n junction structure is to obtain the highest rate of radiative recombination possible. By applying a bias voltage to the junction, the current density is increased, and as it keeps increasing, the quantum well begins to fill up with electrons and holes. The quantum well confines the injected carriers which increases the local recombination rate. If the carrier injection from the electronic source is at higher rate than the recombination rate, then the extra carriers will conduct across the quantum well. To maximise the number of carriers that undergo radiative recombination, Multi-Quantum Well (MQW) structures are used with a repeating series of narrow and wide band gap material [100]. To block the carriers that may still escape the MQW, especially the electrons because of their larger diffusion constant, an Electron Blocking Layer (EBL) is often added. As the EBL layer is made of a different material, it creates an energy barrier in the conduction band which forces the electron to remain in the active region, hence reducing the number of electrons escaping. Fig. 2.3 shows an example of a quantum well and MQW structures with the energy levels and the

EBL layer. For more information on the electrical properties of a semiconductor LED, please refer to [100].

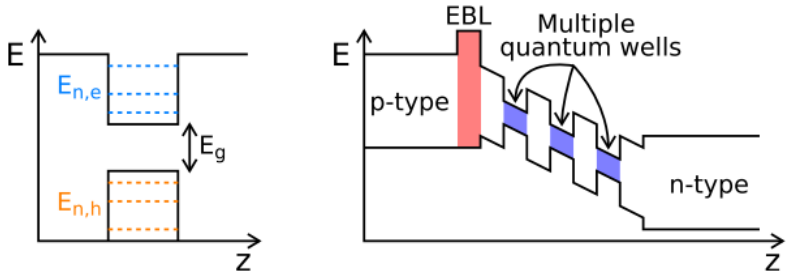


Figure 2.3: Potential well showing discrete electron and hole energy levels (left) and modern LED band diagram, with multiple quantum wells (MQWs) and electron blocking layer (EBL) [11].

2.1.2 Optical properties

Regarding the optical properties of an LED, in an ideal situation, the active region should emit one photon for every electron injected, hence having a quantum efficiency of unity [100]. However, radiative recombination is always in competition with non-radiative processes. The internal quantum efficiency is defined as the number of photons emitted from the active region per second divided by the number of electrons injected into the LED per second. In turn, once the photons are emitted in the active region after a electron-hole pair spontaneous recombination, they should ideally be emitted into free space. However, depending on the material of the LED and various possible loss mechanisms, not all the power is emitted into free-space. This is characterised by the extraction efficiency which divides the number of photons emitted into free space per second by the number of photons emitted from the active region per second. The extraction efficiency can be a severe limitation for high-performance LEDs [100], which can be limited to 50% unless resorting to sophisticated and expensive device processes [98, 102]. Technically, to increase the optical extraction efficiency of an LED, geometry of the device must be optimised and the optical absorption of the epoxy material at the emitted wavelength must be reduced to a minimum. The overall optical extraction efficiency is directly impacted by both the internal quantum efficiency which

directly relies on the LED structure and material of the active region, and by the optical absorption of the epoxy. In general, both the quantum and the extraction efficiency are expressed as the external quantum efficiency expressed as the ratio of the number of photons emitted into free space per second divided by the number of electrons injected into LED per second, in other words it represents the ratio of useful photons to the number of injected charge particles.

To optimise the extraction of the energy generated in the active region of the semiconductor, a parameter called the critical angle for total internal reflection is used to minimise the internal reflection within the semiconductor. The angle of total internal reflection defines the light escape cone of the LED [100]. This angle can be obtained with the well-known Snell's law and relates to the ratio of the air's refractive index to the semiconductor's material refractive index. Light emitted into the cone defined by the angle of total reflection can escape the semiconductor, whereas light emitted outside the cone is subject to total internal reflection. The escape problem is more problematic to obtain a high-efficiency LED as most semiconductors have a high refractive index, hence the acceptable cone will be small and the efficiency low.

In some LEDs, the light extraction efficiency can be enhanced by using dome-shaped encapsulants with a large refractive index which will maximise the emission cone by increasing the angle of total internal reflection [100], see Fig. 2.4. This dome-shape is called an epoxy dome and has a lower refractive index compared to the semiconductor material. In addition, thanks to the dome shape of the epoxy, no internal reflection occurs at the epoxy-air interface. This also means that besides improving the LED external efficiency, the dome can be used as a spherical lens for application requiring directed emission pattern. As an example, the efficiency of a typical semiconductor LED increases by a factor of 2-3 after the encapsulation process, with an epoxy having a refractive index of 1.5.

2.1.3 Modulation rate and applications

The fast modulation capability of LEDs is an important factor in the interest in these devices. To be able to modulate LEDs, the current-injected active region needs to be

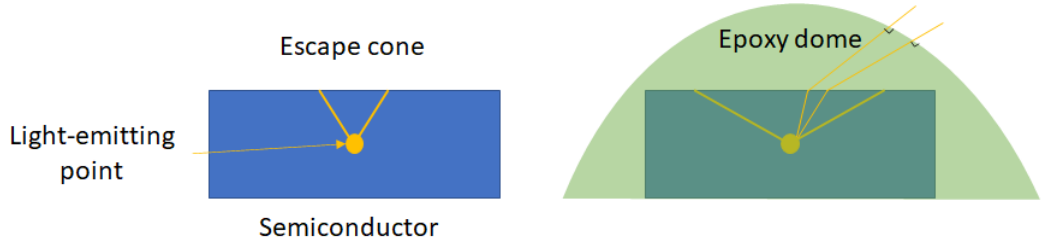


Figure 2.4: LED without (left) and with dome-shaped epoxy encapsulant. A larger escape angle is obtained for the LED with an epoxy dome.

small so that the spontaneous lifetime limits the maximum modulation frequency [100]. While, in comparison, in solid-state lamp applications, the current p-n junction area is large and the modulation frequency is then limited by the diode capacitance. The reduction of the diode capacitance increases the LED modulation bandwidth [100]. In an LED, the spontaneous lifetime of carriers in a direct-band gap semiconductor is of the order of 1-100 ns, which directly depends on the material and the carrier concentration in the active region. In optical communication, LEDs are the most commonly used light source operating from very short, below 1 m, to medium distances, above 5 km [100]. In general, the modulation rates are about tens of Mbit/s but can go up to 1 Gbit/s. To achieve higher modulation rate, micro-LEDs have been developed where the size of the active region area is decreased which in turn reduces the diode capacitance, introduces current density dependent carrier life time effects, and therefore increase the modulation rate. Micro-LEDs are out of scope in this work. Nonetheless, interesting research in visible light communication based on micro-LEDs has been published by the group at the Institute of Photonics, see the references [91, 103–107].

As mentioned above, the light source modulation rate required for the imaging system is about 1 kHz, which is easily achievable with any commercial LEDs. The commercial LEDs that will be used in the thesis are the OSRAM LE RTDUW S2W and the Lumileds LUXEON Z blue colour for the PS and the ToF experiments respectively. The OSRAM LE RTDUW S2W contains four different LEDs, red, blue, green and white. Each LED is a 1 mm square, on a 2x2 chip. The measured electrical-to-optical bandwidth for each white LED is 8.6 MHz. The emitted wavelength for the blue, green

and red LEDs is respectively 450 nm, 520 nm and 632 nm. Only the white LED is used out of the four and a crucial point for the PS experiment is to have a Lambertian emission. The emission angle of this LED has to be wide to avoid obvious divergence effects of the light, which is a requirement for general lighting. The field-of-view of the white LED is equal to 120 degrees. The blue LUXEON LED has similar characteristics in term of the size of the device, although the modulation bandwidth has not been measured.

2.2 SPAD camera

As current CMOS cameras are limited by the level of light that they can detect, mostly due to the size of the pixels, low-level photon imaging has been challenging. High-speed imaging, hence low-light detection, is however needed for different applications such as autonomous driving vehicles, and to enable low-photon capture new sensor technologies are being investigated. Under low-light conditions, photon shot noise dominates the acquisition process and a solution can be found in the Time-Correlated Single-Photon Counting (TCSPC) approach. TCSPC is a photon-efficient statistical sampling technique, whereby photon arrival times are measured relative to a pulsed (laser) source and are recorded in a histogram over many repeated cycles [108]. Many applications such as laser ranging, fluorescence lifetime imaging microscopy and diffuse optical tomography rely on the timing/ low-photon detection properties of single-photon avalanche diodes (SPADs). In this thesis, the TCSPC properties of a SPAD sensor will be used to achieve cm-resolution depth imaging based on an LED-pulsed system. In the following section, the physics behind a SPAD sensor is explained along with a global overview of the current SPAD sensors used in the research field. Next, the LED approach to time-of-flight versus the laser approach will be put into perspective with an explanation on the timing-processing used in Chapter 6.

Over the last decades, one of the main goals of photodetection is to detect single photons over a variety of photon wavelengths and to measure their exact time of arrival at the detector active area [12]. As explained by Durini *et al.* in [12], the hard task of current photodetection systems is to measure as many characteristics as possible of a

single photon. Different methods exist to detect single photons such as Photomultiplier Tube (PMT) [12, 109], superconducting nanowires [110] and silicon photomultipliers when used at low temperature [111]. SPAD sensors offer a number of advantages over these previous methods, which include the spectral response curves that match nicely with visible wavelengths, a good functionality at room temperature, a solid state nature and a small form factor [11, 12, 108]. A full review of the photodetection methods that achieve single photon detection is out-of-scope here and a details can be found in the references cited above.

2.2.1 Physics of SPAD devices

The easiest way to explain the SPAD detector operation is to refer to an Avalanche Photodiode (APD), as a SPAD is engineered so that the avalanche is triggered by a single incident photon. As shown in Fig. 2.5, an APD is a p-n junction whose bias voltage is close to the breakdown voltage, which means that the carrier multiplication will provide high gain [12]. Operating an APD beyond breakdown voltage, i.e in "Geiger mode", creates an electric field strong enough for a single carrier to generate a self-sustaining avalanche [11] and this is the basis of a SPAD. To avoid any damage to the device, the self-sustaining avalanche must be quenched using a passive or an active circuit [112]. Quenching the avalanche will also improve the timing resolution as it will reduce the time when the detector is unable to detect another incoming photon, which is known as the "dead-time". As illustrated in Fig. 2.5, the bias across the detector must be reduced and brought back above the breakdown voltage so that the device is ready to detect another photon. The time-scale of this process will depend on the approach used and parameters of the SPAD and surrounding circuitry. For devices fabricated using CMOS technology, dead times are generally 10 s of nanoseconds [11]. The dead time is an important parameter in photodection as it limits the maximum count rate of the detector. In the event of a photon detection, the SPAD provides a digital signal corresponding to a rising edge pulse with timing accuracy down to the sub-nanosecond level [108, 113].

The spectral response of a SPAD is characterised by the Photon Detection Proba-

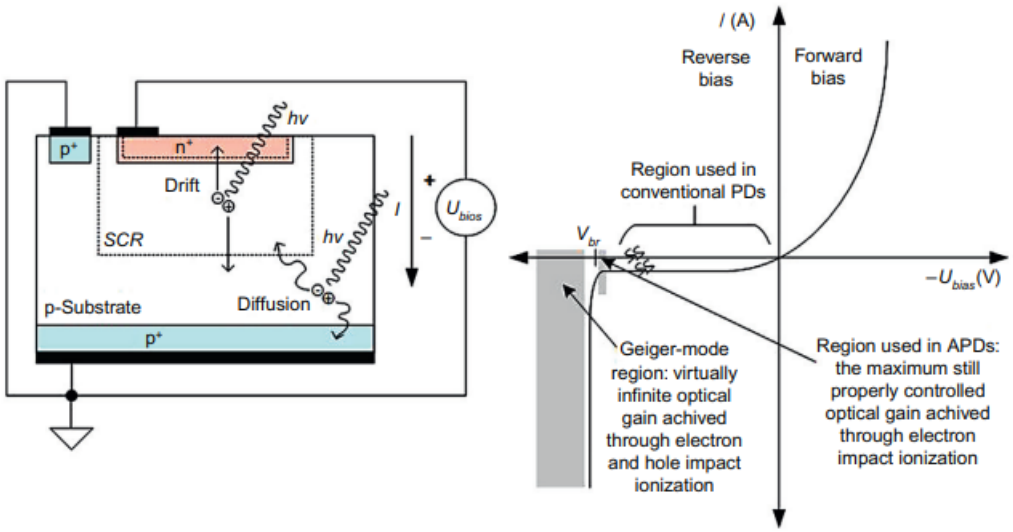


Figure 2.5: Schematic representation of a typical p-n junction (left) fabricated in a CMOS planar technology having a p-type grounded substrate, and a typical I-V characteristic (right) of the same p-n junction, used here as a photodiode, with its different regions of operation: forward bias, reverse bias, with the region without internal gain is used in standard applications, the region slightly below the breakdown voltage is used in APDs, with a minimum controlled (but highly unstable) internal gain caused by impact ionization, and the Geiger-mode region, i.e., the region above the breakdown voltage [12], is used in SPADs.

bility (PDP) or Photon Detection Efficiency (PDE). The PDP is defined as the fraction of incident photons which produce an avalanche and is wavelength dependent due to the wavelength dependency of photon penetration depth into the device [12]. Typical spectral ranges cover the visible and infra-red regions. While the intention of the SPAD is to detect single photons, the avalanche process can be triggered by any charge carriers in the active region, no matter how they are generated [12]. Even with no incident light, there are a number of mechanisms that can lead to unwanted avalanche events from the SPAD, resulting in a Dark Count Rate (DCR). This means that no matter how low the background level is, there will always be a noise component in the output signal. Nonetheless, a good way to decrease the DCR is by the use of the TCSPC technology. In [108], Henderson et al. explain that the window signal, achieved by the Time-to-Digital Converter (TDC) of their SPAD camera, can achieve global electrical masking of photons events, hence, for example, suppressing ambient background in Lidar applications.

2.3 Conclusion

This technical background chapter has covered the physics behind LED and SPAD arrays along with LED properties and optical characteristics. The main thing to remember is the white LED used in this work can be modulated at about 8 MHz which is more than needed as the required modulation frequency will be 960 Hz. LEDs are now becoming ubiquitous in most imaging system as they are cheap and easy to implement. This is why this light source is used as an off-the-shelf device for the 3D imaging work. SPAD arrays are interesting for their timing capabilities and allow to implement a high accuracy pulse tracking and derive high-precision depth map of a scene. They are more and more used in low-photon detection for their high sensitivity. In this work, a SPAD array will be used to implement a ToF imaging setup where the time-correlated single-photon counting will be used to build up a depth map. In the next chapter, a detailed explanation on PS imaging, surface normal integration, Fast Marching will be given along with a full demonstration of the MEB-FDMA modulation scheme.

Chapter 3

Photometric Stereo imaging and MEB-FDMA scheme

Now that the different 3D imaging methods have been introduced, this chapter will detail the Photometric Stereo method and explain the reasons for choosing a calibrated approach. Then, a discussion on the challenge of integrating surface normals is given along with an explanation on the Fast Marching method. To finish, the MEB-FDMA approach is demonstrated along with the decoding algorithm. To easily understand the concept behind the modulation scheme, an explanation on the direct use of the MEB-FDMA will be given. The work in this chapter has been supported by Dr Laurence Broadbent (Aralia Systems) and Dr Johannes Herrnsdorf regarding the photometric stereo algorithm and the MEB-FDMA modulation scheme, respectively. The MEB-FDMA has been developed prior to the start of the thesis and now the aim is to introduce this modulation scheme via a 3D imaging application.

3.1 Calibrated Photometric Stereo imaging

As detailed in chapter 1, photometric stereo is a 3D imaging method that relies on having one fixed camera perspective and different illumination directions to image an object in 3D [39]. An uncalibrated approach would allow for a faster computational system as no lighting calibration is required. However, the uncalibrated PS problem is,

in most cases, ill-posed and brings integration issues within the normal field [114]. For this reason, a calibrated PS method based on 4 images is chosen to estimate the normal field as it is well-posed without resorting to integration [114]. To avoid considering shadow and specularity effects, the surface reflectivity of the imaged object is assumed to be Lambertian [39], *i.e.* the reflection intensity follows a cosine law as a function of angle. In addition, the distance between the LEDs and the object is above the near field lighting area [115] so that the illumination consists of parallel rays from each single source. An orthographic projection is also considered for the formation of the image in the camera's focal plane. The equation of the image formation is expressed below:

$$I_{mg} = A * \mathbf{N}.\mathbf{L} \tag{3.1}$$

where I_{mg} is the measured image, A is the albedo (surface reflectivity), N corresponds to the surface normal components where N_x, N_y and N_z respectively correspond to the horizontal, vertical and depth directions, and $\mathbf{L}$ is the lighting vector of the corresponding LED. Fig. 3.1 illustrates the Eq. 3.1. The LED coordinates are manually measured with respect to a reference point, which corresponds to the centre of the imaged object, and fed into the PS algorithm. No strict calibration process of the LED position is therefore required which speeds up the calculation of the surface normal. Laurence Broadbent developed the presented PS implementation in Python$^{\text{TM}}$ and the computation takes less than a minute to run on a desktop computer.

The first step of the PS algorithm is the determination of the lighting vectors where the LEDs coordinates are first normalised. Then the correct vector length is set by assuming Lambertian emitters of the same intensity:

$$I = I_0 * \frac{xy^2}{d^2} \tag{3.2}$$

$$L = I * \frac{[xyz]}{\|[xyz]\|} \tag{3.3}$$

where $x = [x_1 \quad x_2 \quad x_3 \quad x_4]$, $y = [y_1 \quad y_2 \quad y_3 \quad y_4]$, $z = [z_1 \quad z_2 \quad z_3 \quad z_4]$ are the coordinates of the four LEDs, I is the intensity of the incident LED light at the object

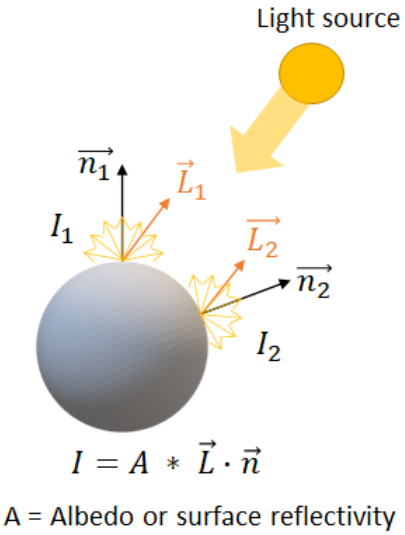


Figure 3.1: Schematic representing a diffuse Lambertian reflection. The diffuse reflected intensity I is a function of the incident light direction **L** and the surface normal **n**. Albedo is the fraction of the incident light source reflected by the surface.

location, d is the distance equal to $d = \sqrt{x^2 + y^2 + z^2}$ and $xy = \sqrt{x^2 + y^2}$. I_0 is the emitter intensity and is set to 1 for each LED. Four images are used as input, one for each LED position, and computed to obtain the albedo A (diffuse reflection of the surface) and the surface normal components **N** using Eq. (3.1):

$$A = \sqrt{\sum (I_{mg}.\mathbf{L})^2} \tag{3.4}$$

$$\mathbf{N} = \frac{I_{mg}.\mathbf{L}}{A} \tag{3.5}$$

The normal field obtained with Eq. (3.5) is then integrated to obtain the surface reconstruction of the imaged object.

3.2 Surface normal integration

Photometric Stereo imaging retrieves the surface normal components of an object, along with the albedo, but does not ultimately provide the surface reconstruction. By

integrating the surface normal components, the 3D shape of the imaged object can be retrieved. To set the problem [114, 116], a usual XY grid is considered where normal vectors are given at each grid point (x, y) as:

$$\mathbf{N}(x, y) = \left(\frac{\delta Z}{\delta x}, \frac{\delta Z}{\delta y}, -1\right) \tag{3.6}$$

where $\mathbf{Z}$ is the height function. By convention [116], the partial derivatives of Z are equal to $\frac{\delta Z}{\delta x} = -\frac{\mathbf{N_x}}{\mathbf{N_z}}$ and $\frac{\delta Z}{\delta y} = -\frac{\mathbf{N_y}}{\mathbf{N_z}}$, where $\mathbf{N_x}$, $\mathbf{N_y}$ and $\mathbf{N_z}$ are the x,y,z components of the surface normal. One way to solve the system of partial differential equations is to minimize the following error function over the entire grid [116]:

$$\varepsilon(Z) = \sum_{i,j} \left(\frac{\delta Z}{\delta x} + \frac{\mathbf{N_x}}{\mathbf{N_z}}\right)^2 + \left(\frac{\delta Z}{\delta y} + \frac{\mathbf{N_y}}{\mathbf{N_z}}\right)^2 \tag{3.7}$$

where $\varepsilon(Z)$ is the quadratic function of the values of Z at the grid point (i, j). Different methods exist to minimize the error function and integrate the surface normal.

3.2.1 State of the art

The integration of surface normal vectors for the computation of a surface in 3D space is a classic problem in computer vision and is needed for shape-from-shading [117, 118] or photometric stereo [39] applications. Since 1986, many computational methods have been developed to solve the problem of surface normal integration. A classic technique based on calculus of variation is given by Horn and Brook [117], which relies on a general Jacobi iteration approach. A faster local solution, based on a direct line-integration scheme and first introduced by Wu and Li [119], demonstrates a computational method that is more efficient than Horn and Brook. However, because of the local nature, the reconstruction solution depends on the integration path, which can lead to an accumulation of errors as local approaches are sensitive to noise and discontinuities. Another famous method is given by Frankot and Chellappa [120] and proposes a frequency domain method to overcome the possible non-integrability of the gradient field. Although this method is more efficient than an iterative approach, it can only apply to non-rectangular computational domains. Improvements based upon these methods have

been investigated since then and are referenced here for more details [121–123].

A recent survey from Queau *et al.* [114] reviews most of the surface normal integration methods and defines useful properties when comparing each method. Unsurprisingly, a basic requirement in 3D-reconstruction is the method's accuracy [114]. Moreover, Queau gives five other properties that an integrator should achieve, namely [114]:

- **Fast** - should be as fast as possible.
- **Robust** - should be robust to noisy field.
- **Free Boundary** - should be able to handle free boundary.
- **Discontinuities** - should preserve the depth discontinuities, *i.e* perform PS on a whole image without segmenting the scene into different parts without discontinuity.
- **Non-rectangular domain** - should be able to work on a non-rectangular model.
- **No Parameters** - should have no parameter to tune.

It is clear that an integrator cannot meet all of those properties at the same time. For example, in their paper on variational methods for normal integration, Queau *et al.* [13] demonstrate that depth discontinuities are hardly compatible with a fast integrator that contains no parameter to tune. To achieve accurate 3D reconstruction of a discontinuity scene [124], the solver will have to adapt to discontinuities and tune at least one parameter which will slow down the computation [13]. It is therefore easier to deal with discontinuities separately and to not include this property in the integrator.

Nonetheless, surface normal integration remains a challenging task as the method needs to be flexible and deal with non-trivial computational domains while achieving high accuracy, robustness and computational efficiency [125]. In the case of PS imaging, artefacts may arise when estimating the normal field based on real-world images, which means that the integrator should be robust to noise and outliers. Moreover, it is well-known that PS imaging does not support discontinuities. Indeed, the sharp gradients representing the transition from the object to image and the background is a difficult

feature to resolve [125]. Image segmentation is therefore required in PS imaging to only integrate the surface normal within the region of interest. For this reason, for the integrator to be efficient and accurate, the computational domain should not be strictly rectangular but adaptable to the region of interest or a defined mask [125]. Another key requirement to consider is the integrator's computational efficiency, as camera technology is constantly evolving and higher resolution images will be acquired [125].

For these different reasons, an integrator than can achieve fast computation of non-rectangular domains is applied. The Fast Marching method is a low computational complexity integrator that can achieve efficient computation of non-trivial domains [116, 125–127]. First employed by Ho *et al.* [116], the approach is based on an analytic formulation of the integration task in terms of an eikonal equation. Galliani *et al.* [126] improved the method by considering a discrete approach which is more accurate and robust than [116]. In some recent work, the Fast Marching method was also found to be useful as an initialisation step to a precondition Krylov subspace method which can achieve higher accuracy [125, 127]. Another advantage of a fast marching integrator is the low memory requirement [125].

3.2.2 Fast Marching Method

First introduced by Sethian *et al.* [128], Fast Marching (FM) is a computational technique used to track propagating surfaces and has proven to be an efficient method to solve the Eikonal equation [116, 128]. The Eikonal equation is a non-linear partial differential equation encountered in problems of wave propagation and provides a link between physical optics and geometric optics [129]. Sethian *et al.* demonstrated that the problem of integrating surface normal vectors is close to the solution of the Eikonal equation [128]. Indeed, by using the same convention as Ho *et al.* [116] and based on the Eq. (3.7), the gradient of the height Z can be expressed as:

$$\|\nabla Z\| = \sqrt{P^2 + Q^2} \tag{3.8}$$

where

$$P = \frac{\mathbf{Nx}}{\mathbf{Nz}} = -\frac{\delta Z}{\delta x} \tag{3.9}$$

$$Q = \frac{\mathbf{Ny}}{\mathbf{Nz}} = -\frac{\delta Z}{\delta y} \tag{3.10}$$

and corresponds to an Eikonal equation which is mathematically expressed as:

$$\|\nabla u(x)\| = \frac{1}{f(x)}, x \in \Omega \tag{3.11}$$

and subject to $u|_{\delta\Omega} = 0$ where Ω is an open set in $\mathbb{R}^n$ with a well-behaved boundary [129]. Physically, a solution $u(x)$ to the Eikonal equation is the shortest time needed to travel from the surface boundary to a point x within the surface with $f(x)$ being the speed at x. In this situation, the right hand side of Eq. (3.8) is positive and according to the definition of Eq. (3.11), and $f(x)$ is a function with positive values. The similitude between Eq. (3.8) and Eq. (3.11) is therefore clear.

When solving the Eikonal equation for Z, the initial point has to be the global minimum. However, because of the noise present in the data and because of local minima, there is no guarantee that the global minima can be easily determined, as shown in Fig. 3.2 a) where a sinusoidal function contains multiple local minima. To deal with this local minimum issue, Ho *et al.* proposed to solve the Eikonal equation for another function W of the form [116]:

$$W = Z + \lambda f \tag{3.12}$$

where Z is the height, λ a constant and f the square Euclidean distance function that is determined as $f = (x - u_x)^2 + (y - u_y)^2$ with (u_x, u_y) the starting point coordinates, which corresponds to the boundary condition that is set manually. The purpose of the function f is to deal with the local minimum value issue of any gradient grid. In fact, f will cancel all the critical points of the grid by creating one global minimum value: (u_x, u_y) [116] as shown in Fig. 3.2 (b). The Eikonal equation for W is equal to:

$$\|\nabla W\| = \sqrt{(P + 2\lambda x)^2 + (Q + 2\lambda y)^2} \tag{3.13}$$

The partial derivative equation (PDE) Eq. (3.13) is a non-linear Hamilton-Jacobi equation of which the solution has to be understood in the framework of viscosity solutions [130]. By understanding Eq. (3.13) as a wave propagation equation and by comparing it to the upwind scheme used in computational fluid dynamics to solve hyperbolic partial equations numerically, Sethian explains that the FM embodies the same properties as the upwind scheme approach used in fluid dynamics [128]. By definition, an upwind scheme numerically simulates the direction of propagation of information in a flow fluid and is based on the discretisation of the hyperbolic PDE by using differencing biased in the direction determined by the sign of the characteristic speed [131, 132]. In other words, the upwind scheme numerically follows the direction of a wave propagating at a determined speed. Fig. 3.2 shows that Eq. (3.13) has an hyperbolic shape, which explain why the FM original design already embodies an upwind strategy. For all these reasons, it makes more sense to apply the same approach as [125, 127] and to use a discrete formulation based on the discrete derivative difference. The upwind discretisation of the partial derivatives for the function f reads as:

$$f_x := \left[\max\left(\frac{f_{i,j} - f_{i-1,j}}{\Delta x}, \frac{f_{i,j} - f_{i+1,j}}{\Delta x}, 0\right)\right]^2 \tag{3.14}$$

and analogously for f_y, where Δx and Δy are the grid widths. By applying the same discretisation to $\|\nabla W\|$, a quadratic equation is obtained and needs to be solved in every grid point except at the boundary point (u_x, u_y).

Based on causality, the upwind scheme [128] updates the gradient from smaller values to larger values in one pass, meeting each grid point just once [125]. In other words, FM updates the smallest value of the gradient grid first to then compute the next smallest value. As shown in Fig. 3.3 (a), the smallest value is the starting point u defined as a 'known' value, and all the other values of the grid are considered to be in the 'far away band'. In order to simply keep track of the integration stage, all the values that are not determined are considered to be in the 'far away band' which technically means that they are not ready to be updated yet. While the point that are closed to the known value are considered to be ready for an update and therefore

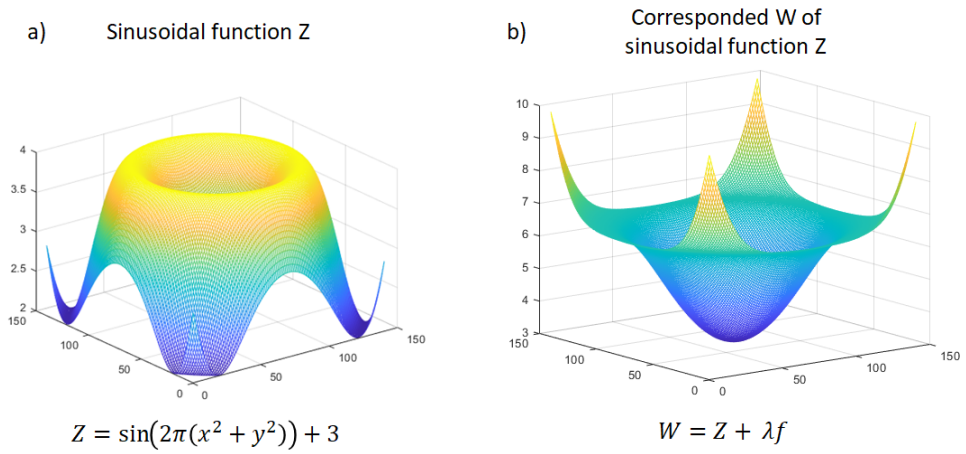


Figure 3.2: a) Graphic of a sinusoidal function that contains multiple critical points, b) Graphic of the corresponding W equation which contains only one global minimum.

are localised as being in the 'narrow band'. At the starting point location, W equals a constant, which is defined as the height of the object, and which sets the boundary condition. The next step consists of adding the neighbours of the known value within the 'narrow band' (Fig. 3.3 b)) and of computing W for each value added, *i.e* ready to be updated. Within the 'narrow band', the smallest value is selected and updated as a "known" value, see Fig. 3.3 c). The neighbours of this new "known" value that are still in the far away band are added to the narrow band where W is computed. The narrow band is computed in a heap-sort fashion [133] to always have the smallest value at the first position in the list to be quickly selected. Steps b), c) and d) are repeated until the narrow band is empty and all values are computed and defined as 'known'. Once W is retrieved then it is easy to recover Z from Eq. (3.12). The FM algorithm in MATLAB^{TM} implementation i based on [116, 125, 127, 128, 133–135] and on Dr. Juan Cardelino's algorithm shared on MATHWORKS^{TM} [136]. By definition, Dr. Juan Cardelino's algorithm computes the distance map to a set of points using the fast marching algorithm and is, however, not generalised to a wider set of applications. In this work, this algorithm is mostly used as a skeleton for implementing the fast marching method with the different stages: 'known' value, 'narrow band' and 'far away band'. As the problem is to integrate the surface normal vectors, N_x, N_y and N_z are the input along with the mask of the object which is used to define a mesh for the

discretisation of the gradient, see Eq. 3.13 and Eq. 3.14. The gradient W is calculated over the grid based on the mesh initially, which is robust to any kind of shape. The borders are therefore not an issue as any configuration is taken into account in the proposed algorithm. The fast marching code implemented is available in the data repository [19] for more details.

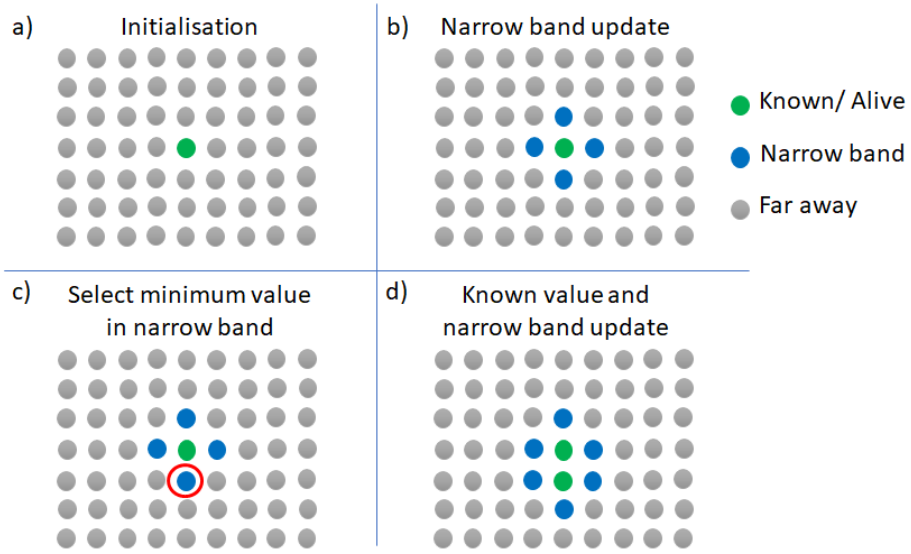


Figure 3.3: Fast Marching propagation process. a) The global minimum is initialised to a constant value and stored as a known value where all the other values are in the far away band. b) the second step is to update the neighbour values into the narrow band and to calculate W, c) the next smallest W value is selected, and d) its neighbours are added to the narrow band where W is computed.

To assess the algorithm, the synthetic results are first compared with Ho *et al.*'s work [116] by plotting the reconstruction of the "Monkey Saddle" function and the STL file of the sphere that was 3D printed (see Chapter 4). Fig. 3.4 shows the superposition of the reconstructed surface on top of the reference data. A perfect match between the two graphs is observed. The only difference implemented in the algorithm presented in this work, compared to Ho's paper, is that the same value of λ is chosen across the gradient grid. In order to select the best value, the Fast Marching algorithm is run for different value of λ and plot the Root Mean Square Error (RMSE) for each value in

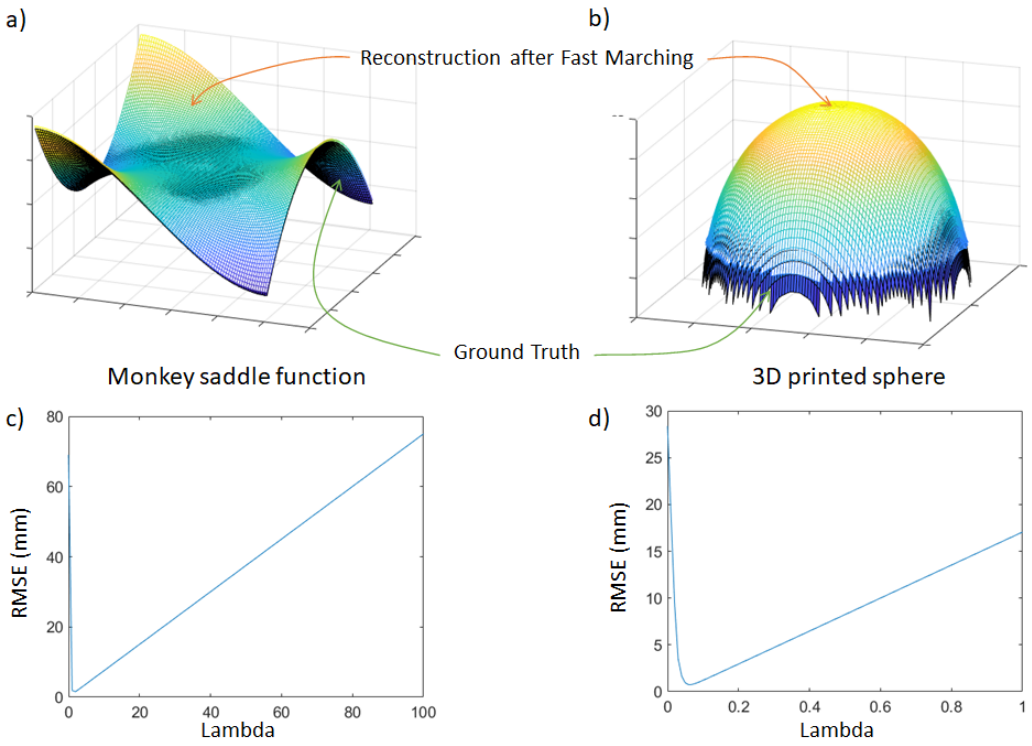


Figure 3.4: Fast Marching 3D reconstruction. a), b) Superposition of the ground truth (black curve) with the FM 3D reconstruction for the monkey saddle function and the 3D printed sphere, respectively. c), d) Graphs plotting the RMSE error regarding λ for the monkey saddle function and the 3D printed sphere, respectively.

Fig. 3.4 (c) and (d):

$$RMSE = \sqrt{(\frac{1}{n}\sum_{i=1}^{n} d_i^2)}, \tag{3.15}$$

where n is the number of data pairs and d_i is the difference between measured values and reference values. The RMSE is smaller for small values of λ as long as λ is not equal to 0. Future manipulation will show that the most common value of λ for real data is $\lambda = 0.06$. Finally, the integrated surface reconstruction is compared with a data set made available by Yvain Queau in [13] which consists of the mask of a vase, the gradient values and the ground truth. The gradients are used as inputs with the Fast Marching algorithm implemented in the thesis and a plot shows the result in Fig. 3.5. By also plotting the ground truth, the figure shows that the reconstruction is a good match. The RMSE between the reconstruction and the ground truth is equal to 1.8 a.u (arbitrary units). By normalizing the RMSE with the height of the vase, the normalised RMSE is equal to 2.5%, which is small. These different comparisons show that the FM implemented in this work is correct and accurate.

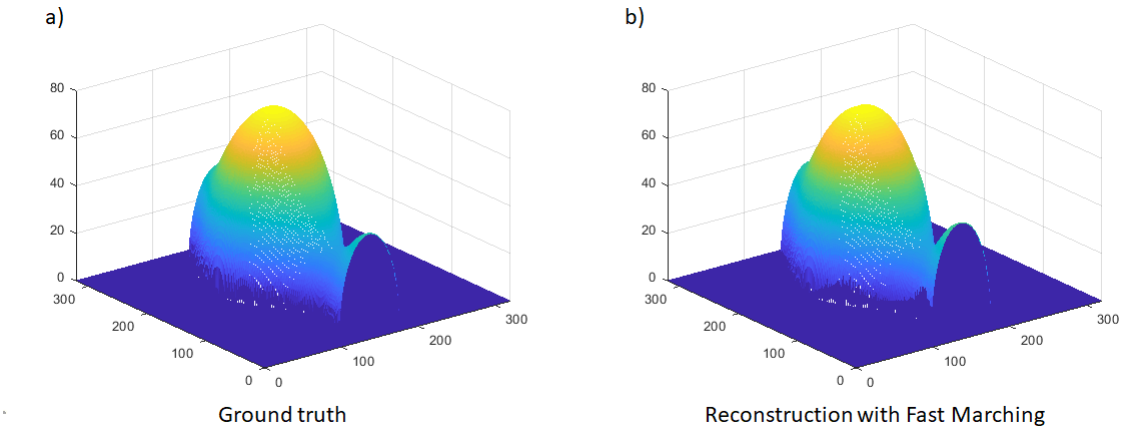


Figure 3.5: Comparison of a vase ground truth from a) Yvain Queau's [13] data set with b) 3D reconstruction using the Fast Marching method.

3.3 Manchester Encoded Binary scheme

A key feature in the experimental work, described in Chapter 4, is the removal of the requirement to synchronize LEDs with the camera or between each other, both of which are needed in the Time Division Multiple Access (TDMA) approach that is used in conventional PS imaging. This is achieved through MA and Frequency Division Multiple Access (FDMA), which has been used by Herrnsdorf *et al.* (in the Institute of Photonics) before to achieve unsynchronized PS imaging [96]. However, FDMA has some drawbacks, in particular strong perceived flicker since some of the LEDs have to be modulated at a fraction of the camera frame rate, and the sinusoidal modulation requires analog control of the LED brightness. To deal with this problem, a bespoke modulation scheme, called Manchester-encoded binary FDMA (MEB-FDMA), has been developed by Dr Johannes Herrnsdorf prior to the start of the 3D imaging project. This works with direct digital modulation of the LEDs, has significantly reduced flicker compared to FDMA, and keeps the advantage of not having to synchronize sources and camera. A comparison of MEB-FDMA with other MA schemes, namely:

- **FDMA** - Frequency division multiple access [96].
- **CDMA** - code division multiple access [137].
- **SDMA** - space division multiple access [138].
- **WDMA** - wavelength division multiple access [139].
- **TDMA** - time division multiple access.

that could be used for PS imaging is provided in table 3.1 where the MEB-FDMA scheme is highlighted in bold. MEB-FDMA and some other MA schemes have the additional feature that they enable visible light positioning of receivers within the imaged scene through a relative signal strength approach (Sec. 3.3.4 in [140]).

In practice, the idea behind the self-synchronised illumination system is to have an easily deployable proof-of-concept that can be retrofitted in public spaces or industrial buildings. The MEB-FDMA scheme is self-clocking which means that no trigger

Table 3.1: **Comparison of MA schemes for N emitters**

MA scheme	FDMA	CDMA	**MEB-FDMA**	SDMA	TDMA	WDMA
Modulation	real-valued	binary OOK	**binary OOK**	binary OOK	binary OOK	none
Duty cycle	50%	50%	**50%**	50%	$1/N$	100%
Frame length	$2N$	$2N$	$\mathbf{2^{N+1}}$	$2\log_2 N$ or $2\sqrt{N}$	N	1 (N fixed to 3)
Computation (flop/frame)	$\sim 10N\log_2 2N$	$4N^2 - N$	$\mathbf{(2^{3N+1}-1)\sum_{i=1}^{N} 2^i}$	0	0	0
Synchronization required LED↔LED	No	Yes	**No**	No	Yes	No
Synchronization required LED↔receiver	No	Yes	**No**	Yes	Yes	No
VLP enabled	3D	3D	**3D**	2D	3D	3D
PS enabled	Yes	Yes	**Yes**	No	Yes	Yes
Visual perception	Poor (flicker)	Good	**Good**	Good	Poor (flicker)	Poor (monochrome luminaires)
Compatible with constant background light	Yes	Yes	**Yes**	Yes	No	No

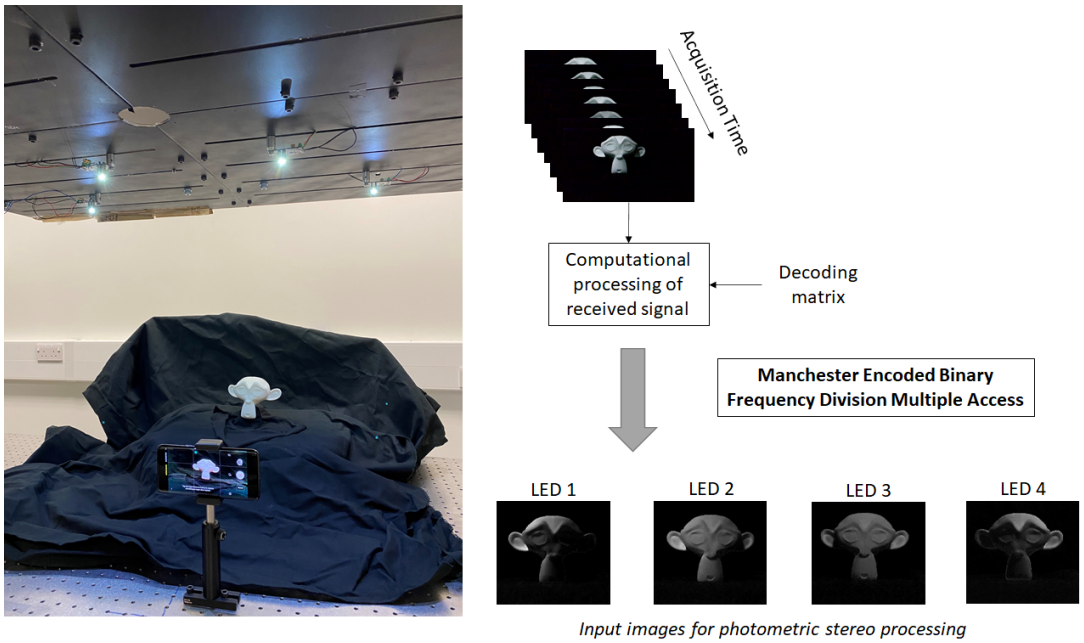


Figure 3.6: Picture showing the experimental setup with the 4 modulated LEDs illuminating the scene and the camera for the frame acquisition (left hand side). To the eye, the illumination is effectively like having the scene fully lit at all times. Schematic representing the stack of frames captured under modulated illumination (top right side), the received images are processed using a pre-determined decoding matrix to separate the 4 effective frames from the perspective of each individual illumination (bottom right side).

is required during the image acquisition. In other words, the camera does not need trigger signals to be sent to the light sources, and each LED has its own orthogonal modulated input signal, which means that the LEDs are not synchronised together. To summarise, the camera and the LEDs are not synchronised so as the LEDs together. As illustrated in Fig. 3.6, the stack of images is captured under modulated illumination. The demodulation process then allows separation of 4 effective frames from the perspective of each individual illumination. Another important feature is that, to the eye, the illumination modulation is fast enough to effectively be like having the scene fully lit at all times. More details about this experiment are given in Chapter 4. This chapter explains the mathematics and the properties of MEB-FDMA. Then, the steps used to calculate the decoding matrix are developed and used in the signal processing of the stack of frames.

3.3.1 Phase invariant Orthogonal Modulation

The important property that allows us to use FDMA without having to synchronize LEDs and camera is that if one frequency carrier experiences an arbitrary phase shift due to the lack of synchronization, it still remains orthogonal to the other frequency carriers. This property is called "phase-invariant orthogonality" and is introduced here formally before describing an alternative modulation scheme that shares this property.

Consider N emitters illuminating the scene over n discrete time steps. Then the $N \times n$ signal matrix $s_{i,j} \in \{-1, 1\}$ describes the time-sequence of on/off states of the individual LEDs. Here, $s_{i,j} = +/-1$ indicates that at time j the i^{th} LED element transmits a binary value of '1'/'0' in On-Off keying (OOK), and after n binary values one 3D image frame is completed. Phase-invariant orthogonality requires that the rows of the matrix s remain orthogonal to each other even if they are time-shifted with respect to each other by an arbitrary phase Δj. Since camera pixels operate as integrating receivers, particularly at high camera frame rates, this requirement can be formalized to Eq. (3.16):

$$\sum_{j=1}^{n} s_{i,j}\left((1-\alpha)s_{i',1+(j-1+k)\%n} + \alpha s_{i',1+(j+k)\%n}\right) = 0$$
$$\forall i \neq i',\ k = 1,\ldots,n,\ \alpha \in [0,1] \tag{3.16}$$

Here the phase shift between rows i and i' is $\Delta j = k + \alpha$, and $\%$ is the modulo operator. Eq. (3.16) represents the requirement from the experimental layout, however, mathematically it is equivalent to the simpler Eq. (3.17):

$$\sum_{j=1}^{n} s_{i,j}s_{i',1+(j-1+k)\%n} = 0$$
$$\forall i \neq i',\ k = 1,\ldots,n \tag{3.17}$$

FDMA with square wave carriers is phase invariant orthogonal. If s is phase-invariant orthogonal, then its Manchester-encoded version $s^{(1)}$ given by Eq. (3.18) - where $\otimes$ is the Kronecker product - is also phase-invariant orthogonal, *i.e.* all the benefits of Manchester encoding are readily available. Matrices of the form $s \otimes \begin{bmatrix}1 & 1 & \cdots & 1\end{bmatrix}$ are also phase-invariant orthogonal.

$$s^{(1)} = s \otimes \begin{bmatrix}-1 & 1\end{bmatrix} \tag{3.18}$$

Decoding of phase-invariant orthogonal encoded signals is less trivial than for CDMA schemes with synchronization. Eq. (3.17) effectively means that the operation of phase-shifting scatters the source signals into orthogonal sub-spaces of $\mathbb{R}^n$. Therefore, to enable successful decoding, the rows of the matrix $s_{i,j}$ need to be complemented by appropriately chosen orthonormal vectors $e^{(i)}_{k,j}$ that together with the rows of $s_{i,j}$ span all of these sub-spaces.

3.3.2 Manchester-encoded binary FDMA

The Manchester code is a form of digital encoding in which data bits are represented by transitions from one logical state to the other, for example the 0 bit would be encoded as 01 and the 1 bit would be encoded as the 10 transition. The Manchester code is therefore twice the length of a common binary sequence however its main advantage is that it can synchronise itself which optimises the reliability and minimises the error rate.

The MEB-FDMA carriers are constructed by starting with binary-valued square-wave FDMA. In order to be phase-invariant orthogonal over a sampling period T, the frequencies ν_i of the square waves must be in a fixed relationship given by Eq. (3.19):

$$\nu_i = \frac{p_i}{T} \qquad p_i \in \mathbb{N}^+ \tag{3.19}$$

A convenient choice of the integer values p_i is given by Eq. (3.20), in which $i = 1, \ldots, N$ identifies each LED:

$$p_i = 2^{i-1} \tag{3.20}$$

This means that the frame length $n^{(0)}$ without Manchester encoding is $n^{(0)} = 2^N$. A binary FDMA emitter signal $s^{(0)}_{i,j}$ is then constructed:

$$s^{(0)}_{i,j} = (-1)^{\lceil j/p_i \rceil}, \quad i = 1, \ldots, N \quad j = 1, \ldots, n^{(0)} \tag{3.21}$$

When using $s^{(0)}$, individual emitters may have long on and off times, leading to unacceptable visual flicker. Therefore, Manchester encoding is used:

$$\begin{aligned} s_{i,j} &= s^{(0)}_{i,j} \otimes \begin{bmatrix} -1 & 1 \end{bmatrix} \\ &= \begin{cases} (-1)^{\lceil j/2^i \rceil}, & j \text{ even} \\ (-1)^{1+\lceil (j+1)/2^i \rceil}, & j \text{ uneven} \end{cases} \end{aligned} \tag{3.22}$$

If the emitters are modulated with $s_{i,j}$ according to Eq. (3.22) using OOK, then they provide MEB-FDMA. A decoding algorithm for MEB-FDMA is provided in sec-

tion 2.4.4 and underlying mathematical proofs are given in the appendix A at the end of the thesis.

3.3.3 Properties of MEB-FDMA

For MEB-FDMA, similar to FDMA, any DC offset can be added to the received signal without affecting the decoding result. This is a prerequisite for applying the scheme to LED illumination since the intensity-modulated LED emission has only positive values. Furthermore, it allows installation of additional lighting fixtures that either do not carry a modulation signal or carry one at a much higher frequency, *e.g.* for LiFi.

Another remarkable property is that the transmitter and receiver can use the same sampling rate. This is surprising because the scheme uses Manchester encoding and the Nyquist theorem requires oversampling by a factor 2 to reliably identify each Manchester encoded bit. However, by requiring the modulation to fulfil the stringent criterion Eq. (3.16), the scheme was implicitly designed such that not every single Manchester bit needs to be identified individually. This property of the scheme means that the frequency of the LED modulation is the same as the camera frame rate and thus flicker is significantly reduced.

The number of OOK-bits needed for a single frame in MEB-FDMA scales exponentially with the number of emitters. Therefore, this modulation scheme is suitable for modest numbers of modulated emitters illuminating the camera field of view, typically 4-6 emitters in the suggested application.

3.3.4 MEB-FDMA Decoding Algorithm

In conventional FDMA, decoding is achieved by a Fourier transform where for each frequency ν_i, there are two real-valued orthogonal base vectors, namely $\sin(2\pi\nu_i t)$ and $\cos(2\pi\nu_i t)$. Note that these two FDMA base vectors are phase shifted by $\pi/2$ with respect to each other. Decoding of an MEB-FDMA signal is analogously done via a set of orthogonal base vectors resulting from phase shifts of the original transmitted $s_{i,j}$. The received signal $r^{(i)}$ from the i^{th} emitter at a given receiver (camera pixel) will have a certain intensity I_i and it will be phase shifted by a phase $\Delta_j = k + \alpha$ compare to

Eq. (3.16):

$$r_j^{(i)} = I_i \left((1-\alpha) s_{i,1+(j-1+k)\%n} + \alpha s_{i,1+(j+k)\%n} \right) \tag{3.23}$$

Eq. (3.23) assumes that the intensity is integrated during the sample time, which is the case in this experiment. The actual received signal r is then the sum of the contributions from all emitters:

$$r_j = \sum_{i=1}^{N} r_j^{(i)} \tag{3.24}$$

The task of the decoding algorithm is to calculate the unknown I_i from the known r.

For a given i an $n \times n$ matrix $S_{k,j}^{(i)}$ is constructed:

$$S_{k,j}^{(i)} = s_{i,1+(j-1+k)\%n} \tag{3.25}$$

where s is given by Eq. (3.22). S does not have full rank:

$$l_i = \text{rank}(S^{(i)}) = 2^i < n \tag{3.26}$$

Its rows span the subspace of $\mathbb{R}^n$ that contains all phase-shifted versions of $s_{i,j}$, but they are not orthogonal. Therefore an orthonormalisation algorithm "GS" is used to construct $l_i \times n$ matrices $D^{(i)}$:

$$D^{(i)} = \text{GS}(S^{(i)}) \tag{3.27}$$

In this work, GS is the stabilised Gram-Schmidt (GS) algorithm [141], but other orthonormalisation procedures can be used as well. The orthogonal decoding matrices $D^{(i)}$ can be used to decompose r into its orthogonal components c_k:

$$c_k^{(i)} = \sum_{j=1}^{n} D_{k,j}^{(i)} r_j \tag{3.28}$$

In conventional FDMA, the signal I_i would be the amplitude of the vector $c^{(i)}$. However, due to the non-orthogonality of the process in Eq. (3.23), a slightly different approach

needs to be taken in MEB-FDMA. In a first step, $r^{(i)}$ is reconstructed:

$$r_j^{(i)} = \sum_{k=1}^{l_i} c_k^{(i)} D_{k,j}^{(i)} \tag{3.29}$$

Then each of the sequences $s_{i,j}$ has at least one occasion where two 1's are transmitted directly after each other. This means that $r^{(i)}$ will have at least one entry with the intensity corresponding to a synchronized 1-level, and this intensity will be the maximum of $r^{(i)}$. Therefore:

$$I_i = \max(r^{(i)}) \tag{3.30}$$

Eqs. (3.25) to (3.30) are the decoding algorithm for MEB-FDMA. The orthonormalisation process in Eq. (3.27) is computationally demanding. However it needs to be carried out only once upon system manufacture/installation and then the matrices $D^{(i)}$ can be stored permanently in a memory on the receiving clients.

3.4 Conclusion

This methodology chapter has covered the relevant computational methods for the PS experimental work focused on the integration of the surface normal with the fast marching method. One main contribution is the adaptation of the fast marching algorithm to the surface reconstruction problem with free-boundary. The discussion of surface normal integration has demonstrated that the fast marching is one of the most efficient, low-complexity methods that can easily adapt to non-trivial computational domains. Calibrated PS imaging has also been discussed. Another main contribution from Dr Johannes Herrnsdorf is the full demonstration of the MEB-FDMA modulation scheme which has two strong properties, namely the synchronisation-free aspect between the camera and the LEDs and between the LEDs themselves along with the flicker-free property. This chapter is a building block for the experimental work which will apply the MEB-FDMA scheme and show that its useful properties are very important in making a deployable synchronisation-free 3D imaging system.

Chapter 4

Top-down illumination Photometric Stereo Imaging

In Chapter 3, the demonstration of the full MEB-FDMA modulation scheme, along with the detailed explanation of the PS imaging process, has been given to introduce this chapter focused on its experimental implementation. Three dimensional reconstruction of objects using a top-down illumination photometric stereo imaging setup and a hand-held mobile phone device is here demonstrated. By employing binary encoded modulation of white light emitting diodes for scene illumination, this method is compatible with standard lighting infrastructure and can be operated without the need for temporal synchronization of light sources and camera. The three dimensional reconstruction is robust to unmodulated background light. An error of 2.69 mm is reported for an object imaged at a distance of 42 cm and dimensions of 48 mm. This work was published in E. Le Francois *et al.*, Optics Express, 29, 1502 (2021) [142], and was also reported as a news release by Optica [143].

4.1 Motivation

In recent years, there has been increasing interest in 3D imaging, as the need to sense and image the surrounding environment has become important in different domains, such as defence and security, robot navigation, autonomous vehicle systems, face recog-

nition and surveillance. Current video surveillance systems in public areas, such as train stations or malls, provide a 2D analysis of a scene by simply recording videos using high-resolution cameras. However, by adding a third dimension, more elements or details would become detectable, and therefore increase the level of security an imaging system can achieve. Reconstructing a 3D scene is challenging, though, as multiple parameters need to be taking into account, such as the acquisition speed of a sensor, light-level with an eye-safety approach, synchronization issues, and computational performance.

Photometric Stereo is one of the most common 3D imaging methods for indoor scenarios. As discussed in previous chapters, PS can achieve better resolution than structured illumination imaging [25,26,36] or state of the art laser scanners [38], offers fast image computation [37], and can deal with objects in motion and untextured areas [1,54]. Compared to stereovision [28,30], only one camera needs to be calibrated, which reduces the computational reconstruction speed, the footprint and the cost [30]. However, the detailed study of PS methods presented in Chapter 1 shows that most PS methods fail to demonstrate an easily deployable imaging technique that could show imaging applications in already existing building infrastructures. If this were achieved, PS imaging could provide an attractive route to using 3D imaging in industrial settings for process control and robot navigation, in public spaces for security and surveillance applications, and for structural monitoring.

There are two major obstacles that inhibit the widespread use of PS imaging for these purposes, which are the compatibility of the PS-specific illumination with indoors or outdoors lighting installations, and the cabling required to synchronize several luminaires with each other and with the camera, which may potentially be mobile. Usually the camera and the luminaires are placed in the same plane, and a particularly common configuration employs four luminaires surrounding the camera in a top/bottom/left/right or X-shaped configuation [4,38,54]. While such a setup is known to provide high-fidelity imaging results, it is incompatible with an application scenario where the luminaires are installed at the ceiling to provide room lighting, and a wall-mounted or mobile camera views the scene from the side, see Fig. 4.1. Current PS systems use cables between the camera and the luminaires to enable synchronization,

which is an undesired complication when retrofitting to existing lighting fixtures. Use of a WiFi or optically encoded clock signal can be a solution to remove the cabling, though additional infrastructure would be needed to implement this and the transmitters, camera and clock signal must be synchronized. Achieving synchronization using a "self-clocking" Manchester-encoded modulation scheme makes the approach described here easy to use, does not require additional infrastructure and can work in environments where WiFi is not available. Finally, traditional PS imaging often has a strong visual flicker and low duty cycle illumination, which is detrimental for indoors or outdoors lighting.

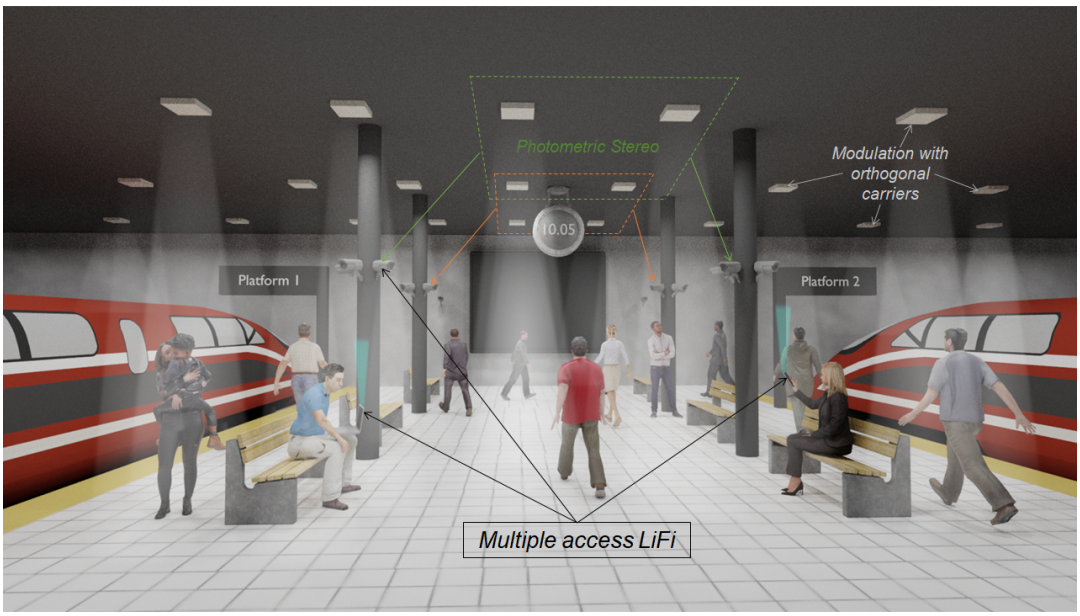


Figure 4.1: Example of a train station representing a multiple access LiFi situation with light communication and video surveillance application using Photometric Stereo imaging.

In this chapter, efforts are focused on making PS imaging synchronization-free, reduce flicker, and demonstrating compatibility with ceiling lighting and both wall-mounted and mobile cameras. In such a scenario, PS imaging would coexist with light fidelity (LiFi) networks [97] or Visible Light Positioning (VLP) [140], potentially using the same light-emitting diode (LED) luminaires for all of these functions [144] as well as general lighting, see the example in a train station in Fig. 4.1. The VLP work

would need an array of LED, for example, which would project different patterns (like fringes) on the scene and would give information on the position of a person or an object. Knowing the position would give more information to then segment the scene and select which light sources to use for the 3D reconstruction. Here, PS imaging is demonstrated using a hand-held mobile phone camera running at 960 fps and ceiling-mounted LEDs, operating in the presence of additional unmodulated lighting. Four LEDs operated by a controller board are mounted on a gantry. These LEDs are modulated with a bespoke binary multiple access (MA) format, referred to as Manchester-encoded binary frequency division multiple access (MEB-FDMA): please refer to Chapter 3 for more details. This modulation scheme removes the need for synchronization and reduces flicker while maintaining a 50% duty cycle. A mobile phone set within the scene acquires a stack of frames at 960 fps, which are then processed to obtain the surface normal components and finally integrated to obtain the topography of the object. As the object is static, the full 3D object is not reconstructed but rather its topography, commonly known as 2.5D reconstruction. For a static 48 mm diameter sphere, a root mean square error (RMSE) of 2.69 mm at a distance of 40 cm is reported, *i.e* a normalised RMSE of 5.6 %, with an angle of reconstruction from the top-down illumination of 120°, which represents 78 % reconstruction of the surface of the sphere.

4.2 Optical acquisition system

As illustrated in Fig. 4.2, the system consists of a mobile phone device (Samsung Galaxy 9), four white LEDs (Osram OSTAR Stage LE RTUDW S2W) placed on a gantry above the object at a height of H = 46 cm, a controller board (Arduino Uno) for the LED modulation and a computer to communicate with the controller board and run the reconstruction program [142]. A series of geometric solids are 3D printed, namely a sphere with a 48 mm diameter, a cube which is 75 mm wide, and a complex shape of a monkey head that is 130 x 94.5 mm^2 wide and 79 mm deep, see Fig. 4.3. In the setup, the geometric centre of the object is the reference (0,0,0) and the location of the LEDs is determined from this reference point. Each LED is modulated with

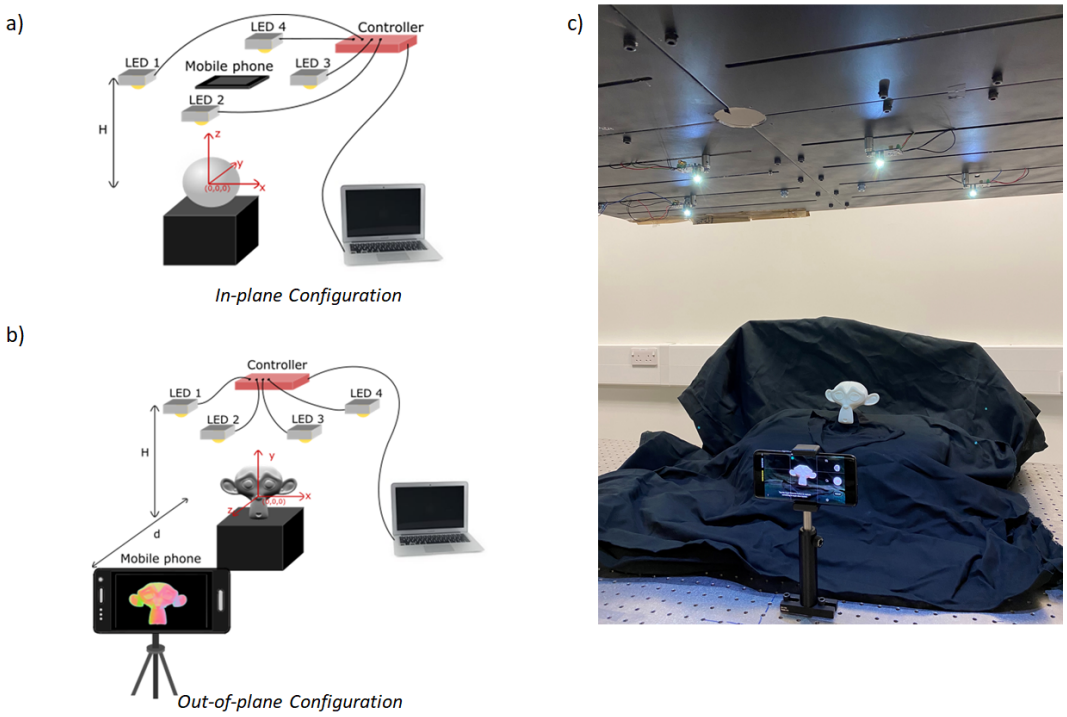


Figure 4.2: Experimental setup. a) Schematic of 'in-plane' photometric stereo imaging configuration, b) schematic of the 'out-of-plane' photometric stereo imaging configuration, c) Picture of the photometric stereo imaging setup in the 'out-of-plane' configuration.

an individual multiple access carrier signal at a frequency of 960 Hz, which is above visual flicker recognition and therefore suitable for digital lighting applications. The carriers are designed such that no synchronization between the LEDs and the mobile phone was required. Each multiple access carrier signal are orthogonal to each other and unsynchronised because of the phase shift experienced by each frequency carrier, see Chapter 3.

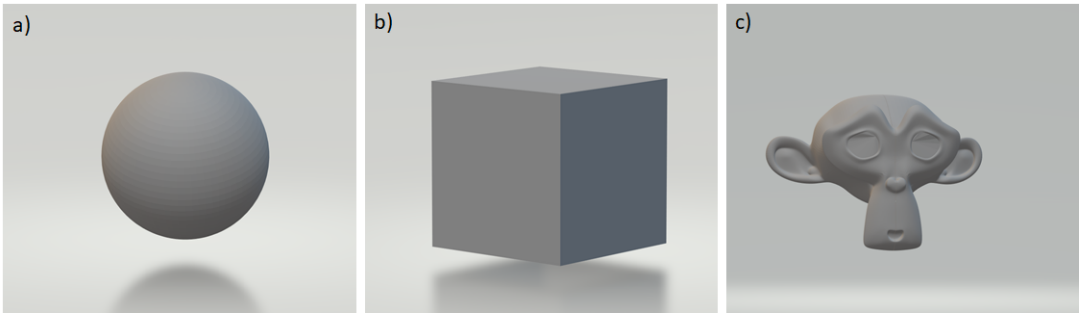


Figure 4.3: Picture of the 3D printed objects, namely a) the sphere (48 mm diameter), b) the cube (75 mm wide) and c) the monkey head (130 x 94.5 mm^2 wide and 79 mm deep).

To benchmark the work, the experimental setup was first set up in an 'in-plane' configuration to reproduce 'conventional' PS imaging [4,38], see Fig. 4.2 a). The mobile phone is placed directly on the gantry with the LEDs surrounding the phone in a left-right-top-bottom fashion, each at a distance of 10 cm from the phone. The top-down illumination setup is then modified to the 'out-of-plane' configuration, see Fig. 4.2 b). The phone is moved from the gantry and mounted on a tripod so that the LEDs and the phone are located in two different planes. The phone is in front of the object on the z axis at a distance of d = 42 cm from (0,0,0) with a field of view of 43 degrees. The LEDs are moved to the following locations (x,y,z): LED1 (-27, 42, 10), LED2 (-14, 42, 35), LED3 (14, 42, 35) and LED4 (27, 42, 10), all in cm.

In both situations, the phone captures frames with a resolution of 1280 x 720 pixels at a rate of 960 fps for 0.2 s, with a black background to simplify image processing. The capture time is limited by the on-device data storage capacity. The Samsung Galaxy 9 can achieve a frame rate of 960 fps thanks to its new triple-stacked camera module which includes a CMOS sensor on top with a readout circuit underneath that

is running four times faster than the previous chip. In addition, the module includes a dynamic random-access memory chip which save the frames and allow a smooth readout. The transmitted signal is encoded as a clock signal, hence no trigger is needed to start the acquisition. As explained in Chapter 3, the LEDs have a 50 % duty-cycle and thanks to the known modulation code fingerprint of each LED, a decoding matrix can be created: please refer to Chapter 3 to see the decoding derived in detail. Fig. 4.4 illustrates an example of the frequency waveforms applied to the LEDs, each waveform is orthogonal to each other and the LEDs are not synchronised to one another. During the transmission, each LED waveform experiences a phase-shift that is taken into account when creating the decoding matrix, so each LED contribution is easily recoverable thanks to their specific fingerprints (a detailed example is given in Appendix A). The received stack of images are demodulated using the decoding matrix on each pixel of each image. To increase the decoding time, it is possible to select less frames to decode from the output video which will impact the quality of the four output images. The lower number of frames required to obtain a good quality image is 32 frames and they are selected from the beginning of the video. The full stack of images (128 frames) is recommended to achieve a higher quality and improve the determination of the surface normals which will then impact the quality of the 2.5D reconstruction. At the end of the demodulation process, four images corresponding to the four different illumination directions are retrieved. The four images retrieved are then processed using established methods [4,39,145] to obtain the surface normal components N_x, N_y, N_z and the albedo A under the assumption of a Lambertian surface: please refer to Chapter 3 for detailed explanations. The last step of the reconstruction program is the integration of the surface normal vectors to obtain the topography of the object using a Fast Marching method [116]. The reconstruction process takes about 3 minutes to run on a desktop PC.

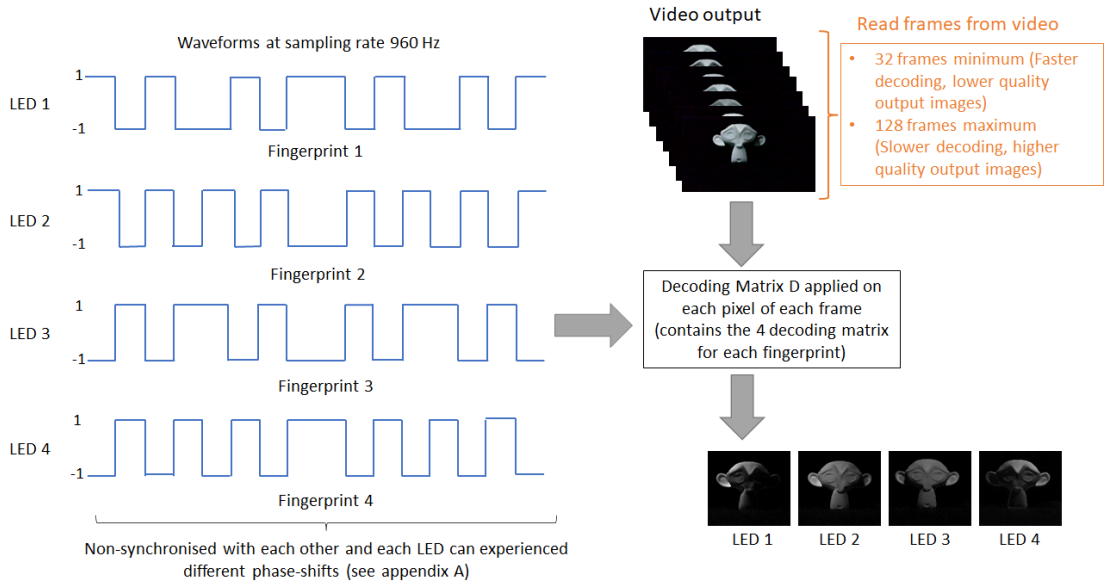


Figure 4.4: Schematic representing an example of possible waveforms applied to each LED. A decoding matrix is calculated for each waveform and concatenated into one main decoding matrix. There is an option to select the number of frames to be decoded (32, 64 or 128 frames) which will impact the quality of the decoded images and the decoding time. Each selected frames are decoded pixel by pixel by the matrix D to then retrieve the four images, one for each LED.

4.3 Results

4.3.1 Proof-of-concept: 'in-plane' configuration

In order to benchmark the work with "conventional" PS imaging [4,38], the 3D imaging system is first evaluated in an 'in-plane' configuration using a spherical test object, see Fig. 4.2 a). Figs. 4.5 a), b), c), d) show the four images obtained after decoding the frames. The reflectivity of each decoded image clearly demonstrates the different illumination direction - left/bottom/right/top respectively - for LED1, LED2, LED3 and LED4. According to the image scale, the amount of light for each direction is similar which means that the decoding process is well-adapted to the frequency.

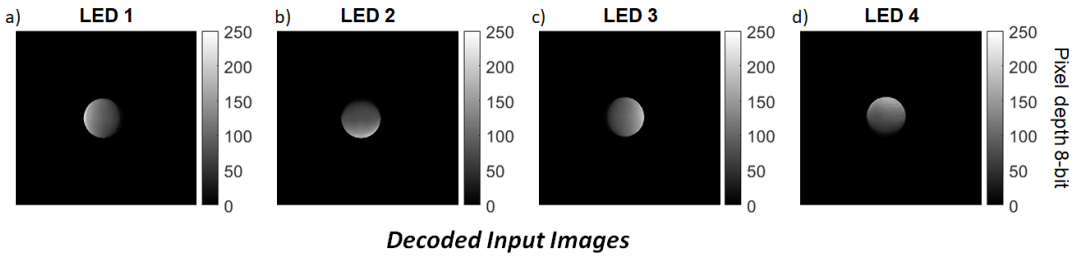


Figure 4.5: Decoded images of the sphere in the 'in-plane' configuration. Obtained after demodulation of the recorded frames for a) LED1, b) LED2, c) LED3 and d) LED4.

Fig. 4.6 displays the surface normal components' ground truth of the sphere along with the surface ground truth. Fig. 4.7 a), b), c) shows the corresponding surface normal components N_x, N_y, N_z obtained with the photometric stereo processing. As expected from the 'in-plane' acquisition scheme in Fig. 4.2 a), N_x correctly distinguishes left and right facing surfaces of the object. Similarly, N_y correctly identifies up and down facing surfaces. Finally, as the back of the object cannot be seen, N_z is always positive with some variations due to the depth of the object. By comparing it with the surface normals of the ground truth of the sphere in Fig. 4.6 a), b), c), the match between N_x, N_y and N_z is satisfactory with some error in N_z. Indeed, N_z decreases slightly faster than the ground truth from its central maximum.

Two 2.5D reconstruction views and the RMSE error map are plotted in Fig. 4.7 d), e), f)

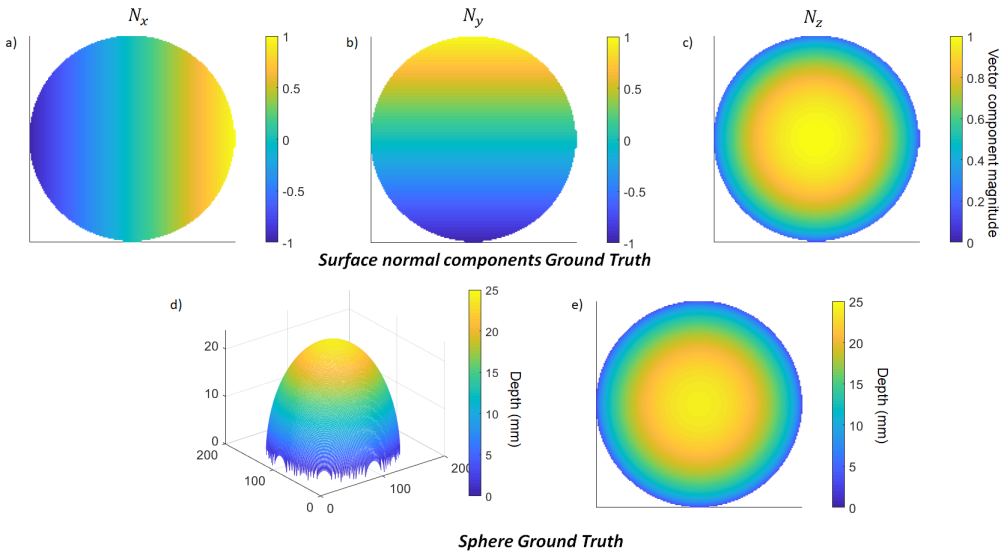


Figure 4.6: In-plane configuration Ground Truth. a), b) and c) plot the surface normal components, respectively Nx, Ny, Nz. d) and e) plot the ground truth surface of the spherical test object, respectively the perspective view and the top view.

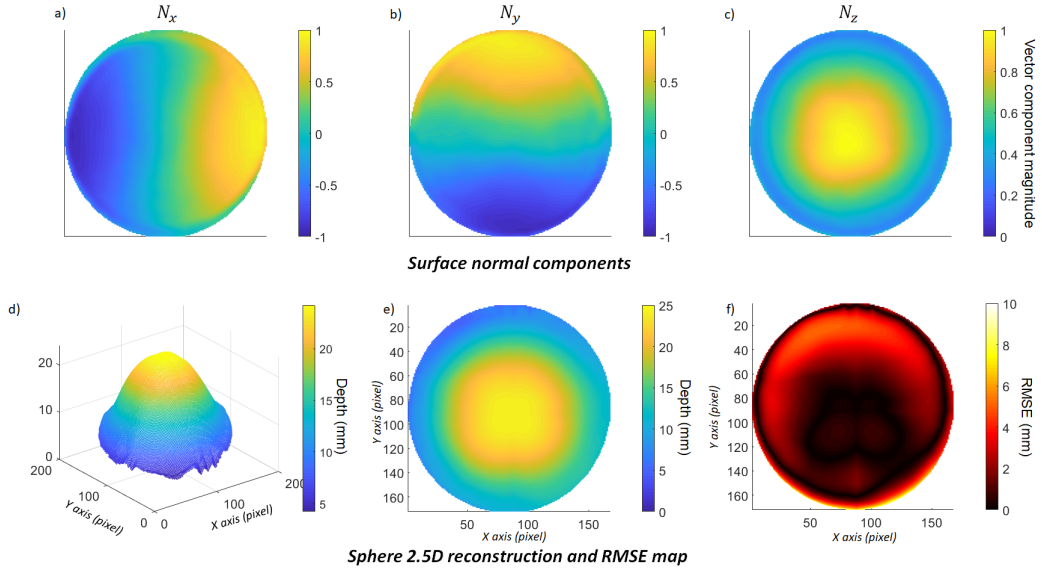


Figure 4.7: 'In-plane' configuration results. a), b) and c) plot the surface normal components obtained after running the photometric stereo algorithm, respectively Nx, Ny, Nz. d), e) and f) plot the 2.5D reconstruction of the spherical test object, respectively the perspective view, the top view and the RMSE error map.

respectively. Following the analysis of [4], the standard deviation is determined by the root mean square error (RMSE) and the Normalised Root Mean Square Error (NRMSE) which are defined as [4]:

$$RMSE = \sqrt{(\frac{1}{n}\sum_{i=1}^{n} d_i^2)}, \tag{4.1}$$

$$NRMSE = \frac{RMSE}{x_{max} - x_{min}}, \tag{4.2}$$

where n is the number of data pairs, d_i is the difference between measured values and reference values, and $(x_{max} - x_{min})$ is the range of measured values. The value of NRMSE is expressed as a percentage, where a lower value indicates less variance, hence a higher accuracy [4]. For this 48 mm diameter sphere, an RMSE of 2.4 mm and a NRMSE of 5 % is obtained. These two results are within the same range of error as in [4]. Fig. 4.7 f) shows the RMSE map over the sphere in order to have a visualisation of the error distribution. The error is higher at the edge of the sphere compared to the middle area, and this can be explained by the Fast Marching method as the error builds up as the reconstruction progresses through the surface. By comparing both the error map with the surface normal components from Fig. 4.7 a), b), c), the error in the 2.5D reconstruction is also closely linked to the error in N_z. By comparing the 2.5D reconstruction and the ground truth plotted in Fig. 4.6 d), the reconstruction matches the ground truth from its central maximum down to 8 mm, which is satisfactory.

4.3.2 Out-of-plane configuration

Section 3.3.1 proves that the 3D imaging system, along with the synchronisation-free modulation scheme, is adapted and provides similar results to other PS work when configured in a conventional way, *i.e* the 'in-plane' configuration. Now, this section investigates the 2.5D reconstruction of the sphere, the cube and the monkey head under top-down illumination, *i.e.* the 'out-of-plane' configuration, see Fig. 4.2 b).

LED images decoded

The decoded images of the three objects are displayed in Fig. 4.8. To assess the reconstruction of sharp edges, the corner of the cube is also imaged. For the four cases, the light level clearly shows the different illumination directions. The brightness is slightly different depending the position of the LEDs. This can be explained by the possibly imperfect match between the camera integration time and the MEB-FDMA scheme, i.e. the integration time might be shorter than the frame duration. Nonetheless, the final four output images of each object are satisfactory to be used in the determination of the surface normal components in the PS stage.

Surface normal vectors

From this set of images, surface normal components (N_x, N_y, N_z) and the albedo were calculated and are displayed in Fig. 4.9. For N_x, left and right facing surfaces are correctly distinguished as vector components with magnitude ranging from -1 to 1. However, Fig. 4.9 e) and f) show that the cube face is not easily distinguished for both left-right and up-down facing surfaces. Indeed, both vector component magnitudes are equal to 0, which is consistent with the vector theory of plane surfaces. Some artefacts can be observed on the edge of the cube. Similarly, N_y indicates up and down facing surfaces correctly, albeit with lower fidelity, and its value range is limited to -0.2 to 1 instead of -1 to 1. The poorer fidelity of N_y is due to the top-down illumination design as the bottom of each object is not suitably illuminated, an effect which is also visible in the albedo plot. Moreover, as the camera is facing the object, N_z is positive and ranges from 0 to 1 with some variations due to the depth of the object. The albedo is normalized and is useful in understanding imperfections in the reconstruction. The albedo is more directional for the sphere and the cube corner than for the monkey, which is caused by the slight brightness variations seen in Fig. 4.8. Importantly, the surface normal components, which are the basis of the topography reconstruction, are observed to be less susceptible to these brightness variations than the albedo.

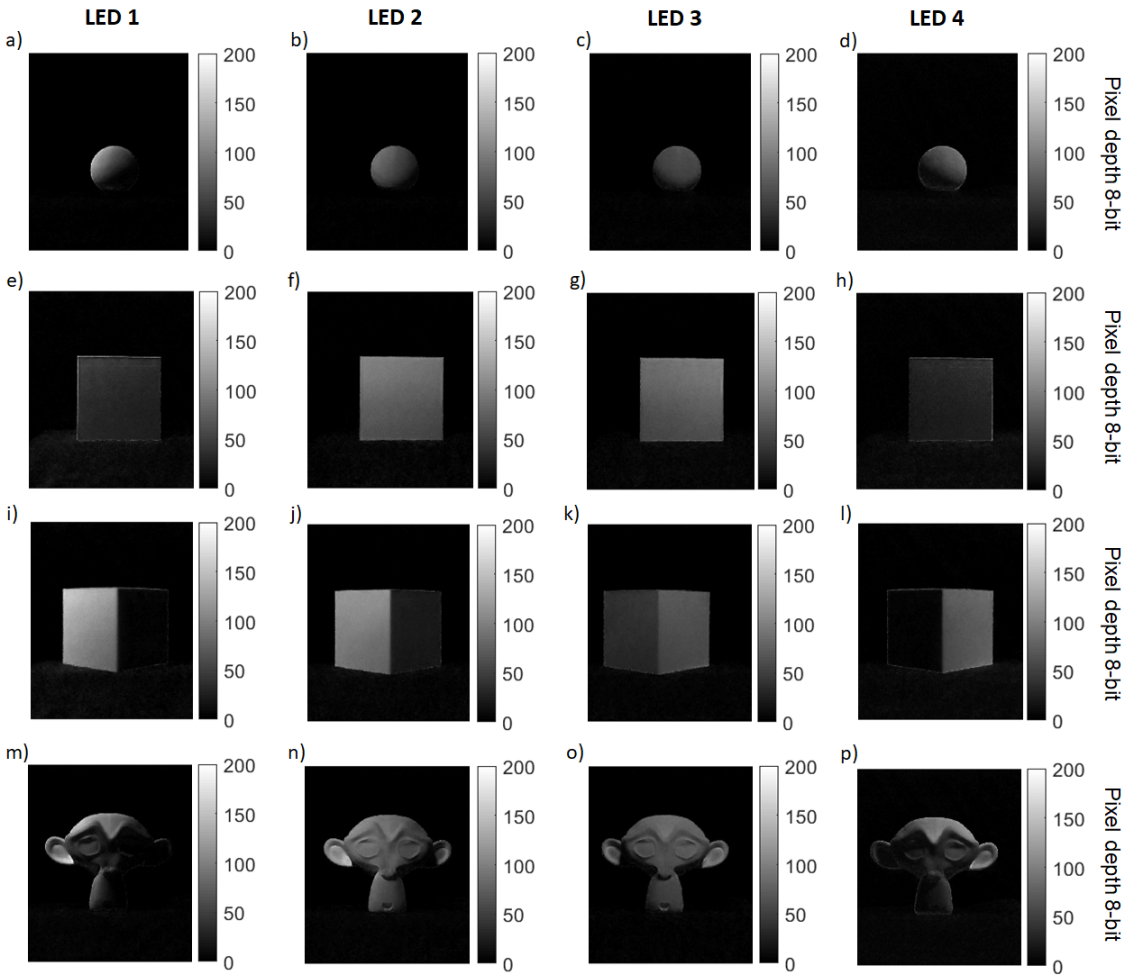


Figure 4.8: Out-of-plane illumination decoded images obtained after demodulation of the recorded frames for LED1, LED2, LED3 and LED4: a), b), c), d) for the sphere, e), f), g), h) for the cube side, i), j), k), l) for the cube corner and m), n), o), p) for the monkey head.

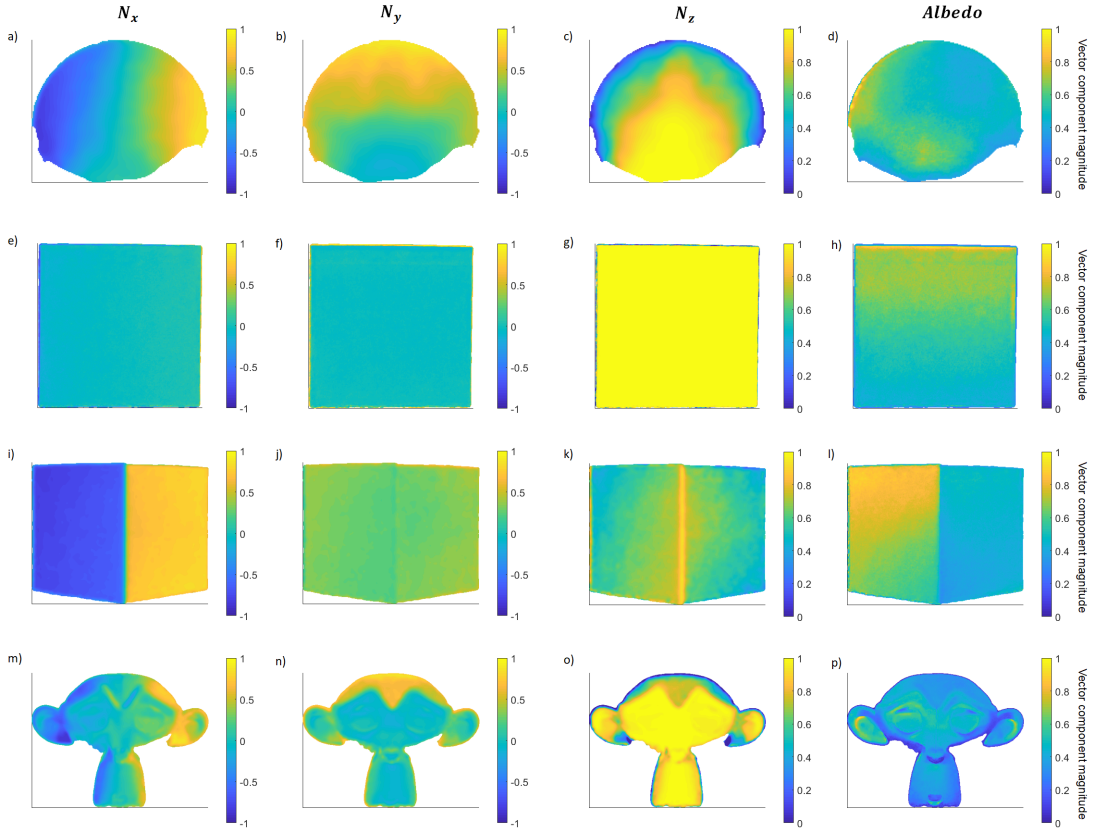


Figure 4.9: Surface normal components and albedo. Obtained after running the photometric stereo algorithm, respectively Nx, Ny, Nz and Albedo: a), b), c), d) for the sphere, e), f), g), h) for the cube side, i), j), k), l) for the cube corner and m), n), o), p) for the monkey head.

Surface Reconstruction

Fig. 4.10, Fig. 4.12, Fig. 4.13 and Fig. 4.14 respectively plot the 2.5D reconstruction of the sphere, the cube side, the cube corner and the monkey head in different perspective views as well as a rendered view using a 3D creation rendering software Blender [146]. Blender is a free, open source, ray tracing based rendering software which can support both CPU and GPU. To render the reconstruction on Blender, a camera is set at a distance of 10 cm from the imported 2.5D reconstruction. The RMSE error map is also plotted for each object.

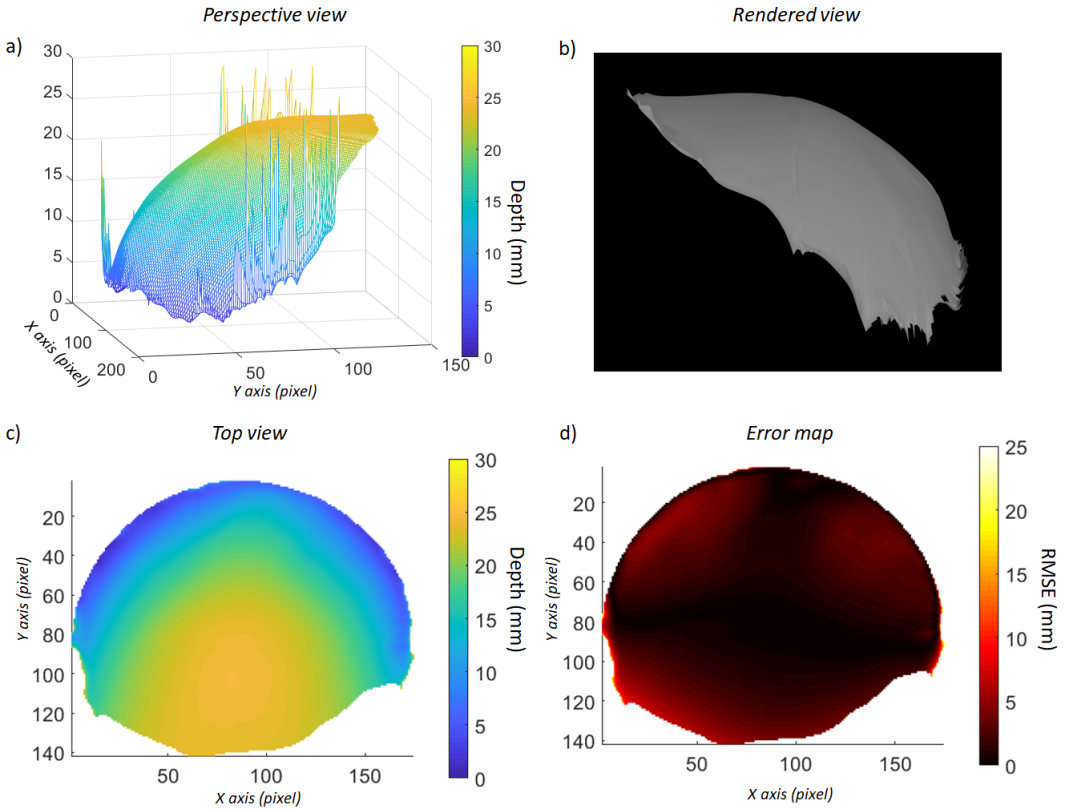


Figure 4.10: 2.5D reconstruction of the sphere, a) Perspective view, b) Rendered view, c) Top view, d) RMSE error map.

For the sphere, Fig. 4.10 a), b), c) show a satisfactory global reconstruction for the top half of the object. The bottom half is poorly reconstructed and is "flat", which is also clearly shown on the rendered view. Because of the lack of information on the negative (downward facing) y axis, 78.4 % of the visible surface is reconstructed.

According to the RMSE error map in Fig. 4.10 d), the most significant error is found at the bottom and on the edge of the sphere, while smaller errors in the top area are related to inaccuracies in N_z. Nonetheless, most of the error stays below 5 mm. Fig. 4.11 shows the projected angle that can be retrieved using the top-down illumination setup where an angle of 120° is retrieved. An RMSE of 2.69 mm and an NRMSE of 5.61 % are obtained within the 78.4 % of the surface reconstructed. Despite the lack of information on downward facing facets, both standard deviation errors for the sphere are within the same range as in [4].

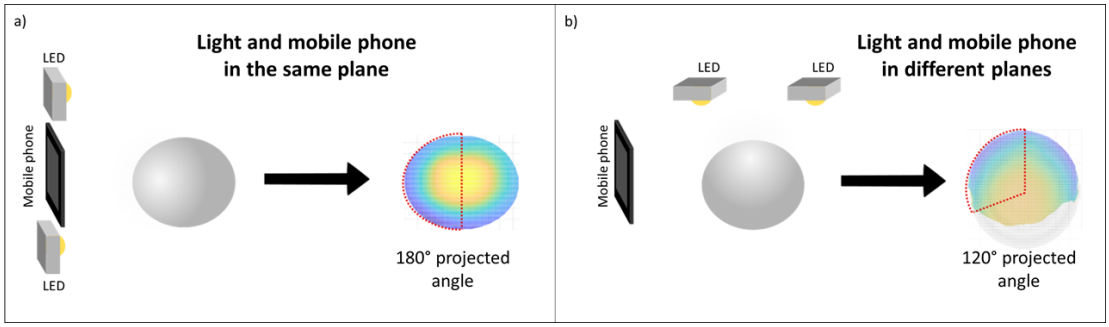


Figure 4.11: Schematic of the difference in the projected angle reconstruction between a) the 'in-plane' configuration and b) the 'out-of-plane' configuration.

For the cube side in Fig. 4.12 a), b), c), the reconstruction of the 'flat' face is slightly tilted due to the asymmetric lighting geometry, though the cube face is entirely reconstructed. In addition the rendered view from Blender, in Fig. 4.12 b), shows a flat surface with no tilt noticeable. According to the RMSE error map in Fig. 4.12 d), the highest error is located on two corners, the yellow and blue corner in the perspective view. Nonetheless, the axis scale is small and ranges from 0 to 3.5 mm. The surface tilt can also be explained by the build up of the error as the Fast Marching algorithm propagates across the object. Using Eq. (4.1) and Eq. (4.2), an RMSE of 1 mm and an NRMSE of 1.36 % are obtained and prove that the tilt can be negligible in terms of the error scaling across the cube. The reconstruction of the cube face is therefore satisfactory and this proves that flat surfaces can also be reconstructed using a top-down illumination approach.

For the cube corner, despite the unequal partition of light on the cube and the

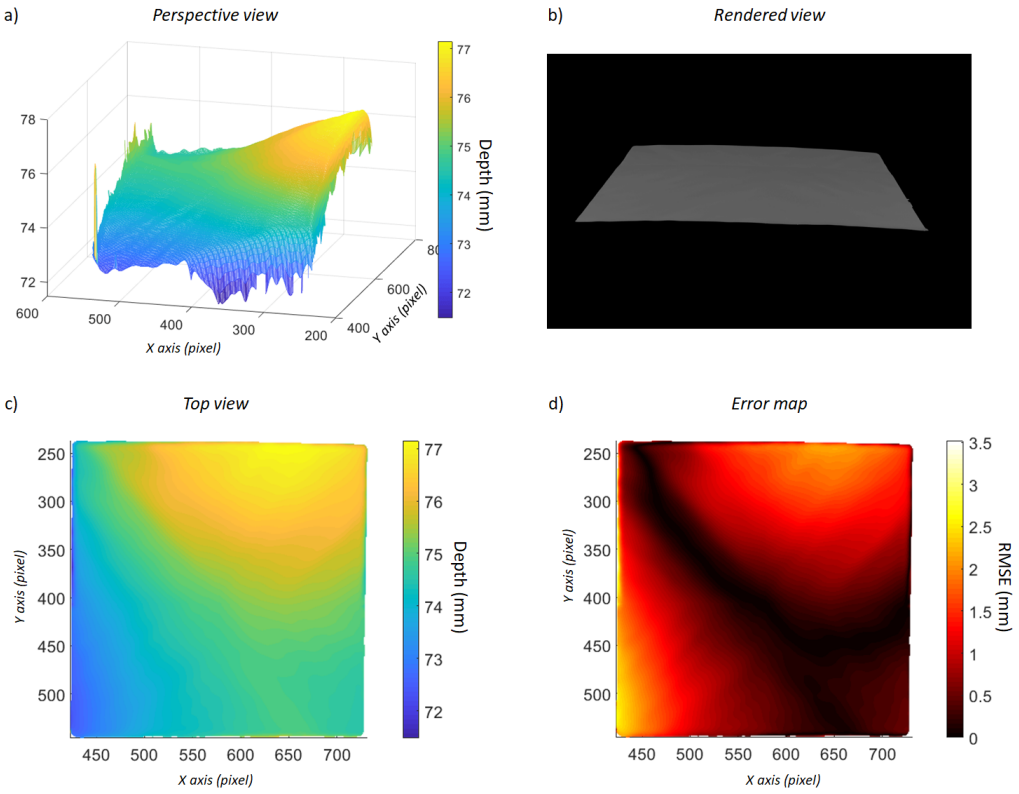


Figure 4.12: 2.5D reconstruction of the cube side, a) Perspective view, b) Rendered view, c) Top view, d) RMSE error map.

important gradient variation, the reconstruction can retrieve the shape of the corner. Nonetheless, the top view in Fig. 4.13 shows that the reconstruction on the edge is deteriorating faster at the bottom of the object. This is explained by the top-down illumination configuration. Moreover, Fig. 4.13 d) plots the RMSE error map where the highest error is reported at the bottom of the cube. The calculated RMSE is equal to 15.3 mm with a NRMSE of 16.6 %, which represents the global error obtained on the error map. Comparing to the two previous objects, this error is higher and can be explained by the angle of the reconstructed corner of the cube. Indeed, the reconstructed cube corner angle is slightly higher than 90°, hence the match between the ground truth and the reconstruction to calculate the RMSE is not perfect although it is well aligned. The error map shows a consistent error across most of the cube which means that despite the error on the angle the global shape is well-reconstructed. Overall, by comparing the reconstruction of Figs. 4.13 a), b), c) with the images of Fig. 4.8 a good match is obtained. On the rendered view, an indent is observed and can be explained by the position of the starting point of the Fast Marching reconstruction process, from which the gradient values are integrated in a propagating fashion. At points with a strong gradient variation, the propagation will happen with very different gradient values in different directions and this can create an indent on the row or the column of the starting point.

Finally, the monkey head has been chosen for its complex features, such as the eyes, the nose and the top of the head. The discontinuity between the face and the ears is a challenge for the Fast Marching algorithm to deal with. Indeed, the 2.5D reconstruction in Figs. 4.14 a), b), c) shows the shape of the nose, the eyes and also the upper head. However, the shape of the ears is harder to determine and the depth is substantially decreasing, which demonstrates the difficulty of the algorithm for dealing with those discontinuities. Using Eq. 4.1 and Eq. 4.2, the RMSE error map is calculated and plotted in Fig. 4.14 d), where an RMSE and an NRMSE of 10.9 mm and 18.9 % are obtained respectively. As expected, the relative position between discontinuous region is not handled well as the highest error is obtained in the ear area and reaches 30 mm. However, features within each region are reproduced with good fidelity, such as the ears,

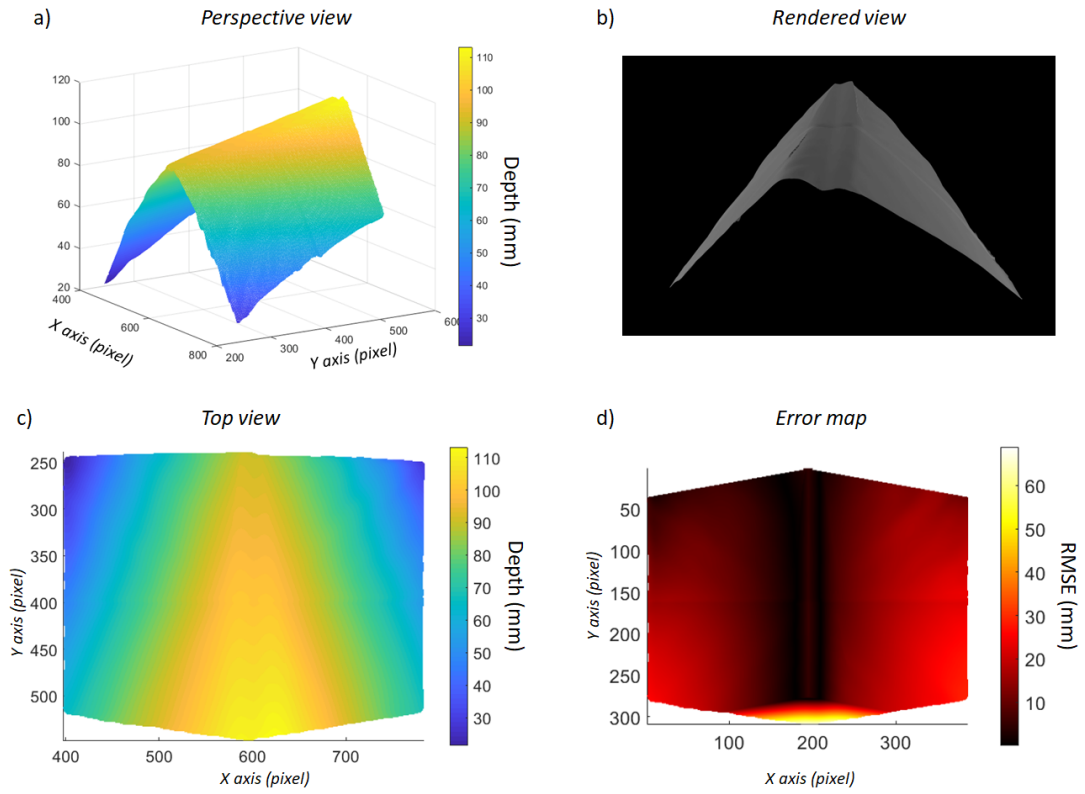


Figure 4.13: 2.5D reconstruction of the cube, a) Perspective view, b) Rendered view, c) Top view, d) RMSE error map.

on the rendered view. In addition, when the error is only determined across the face of the monkey without selecting the ears, then the RMSE and NRMSE are respectively equal to 5.4 mm and 6.7 %. Small depth details, within mm, such as the earlobes and the eyes, are detectable and well reconstructed. The error map shows that most of the error is below 10 mm which is satisfactory for a complex shape object. Moreover, the distance between the face and the ears is about 50 mm and on the top view of the reconstruction the distance between the two features is also about 50 mm.

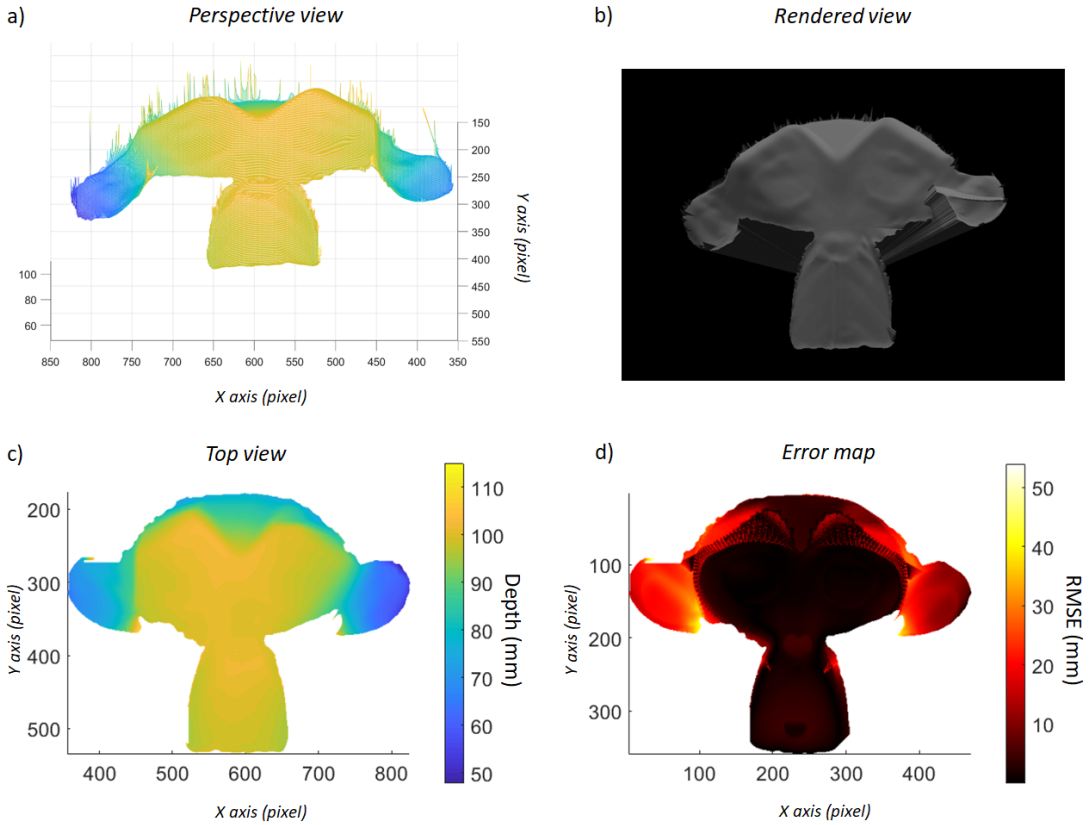


Figure 4.14: 2.5D reconstruction of the monkey head, a) Perspective view, b) Rendered view, c) Top view, d) RMSE error map.

The 3D reconstruction relies on the surface normals, therefore most of the error on the reconstructed object topography will be dominated by the error on the surface normal vectors. Whenever N_z values are close to zero, the gradient integration during the Fast Marching process is numerically ill-conditioned. This phenomenon is clearly shown by the artefacts on the different reconstructions in Fig. 4.10.

4.3.3 Signal to noise ratio

Our modulation scheme can operate in the presence of additional unmodulated lighting. In order to assess its robustness, the reconstruction of the sphere, the cube and the monkey is tested with different levels of background light in the room. For this experiment, the ceiling light of the room is illuminated and the brightness of each LED is controlled by an Arduino board, which modifies the signal power. As the white commercial LED spectrum is broad, a representative measurement of the optical power is measured with a simple photocurrent detector set at one wavelength. By weighting the measured photocurrent by the emission of the white LED at the given wavelength, a global representative measurement of the Signal to Noise Ratio (SNR) is obtained. After measuring the representative optical power of the signal and background light at the object location, the SNR, in dB, is determined following Eq. 4.3:

$$SNR = 10\log_{10}\left(\frac{P_{signal}}{P_{noise}}\right) \tag{4.3}$$

with P_{signal} being the representative optical power of the LEDs at one wavelength and P_{noise} the representative optical power of the ceiling light at the same wavelength.

Fig. 4.15 plots the graphs of the RMSE and the percentage of surface reconstructed versus the SNR for the sphere, the cube face, the cube corner and the monkey head in the 'out-of-plane' configuration. In each case, the SNR ranges from -1 dB to 5.5 dB.

For the sphere, the RMSE does not show a specific trend across the SNR and the error ranges from 4.5 mm to 7 mm which is acceptable for these applications. However, the percentage of surface reconstruction that is achieved over the measured range of SNR does show a dependence on the SNR. Fig. 4.16 shows that as the SNR decreases, the reconstruction of the bottom of the sphere is more and more challenging. By comparing this to the three other graphs, it can be deduced that this dependence is a direct consequence of the shape of the object, *i.e.* the sphere, combined with the top-down illumination configuration. This is explained by the Fast Marching integration process: please see Chapter 3. If at a particular grid point the z gradient is abnormally high, then the reconstructed value will result in an artefact which does not correspond to

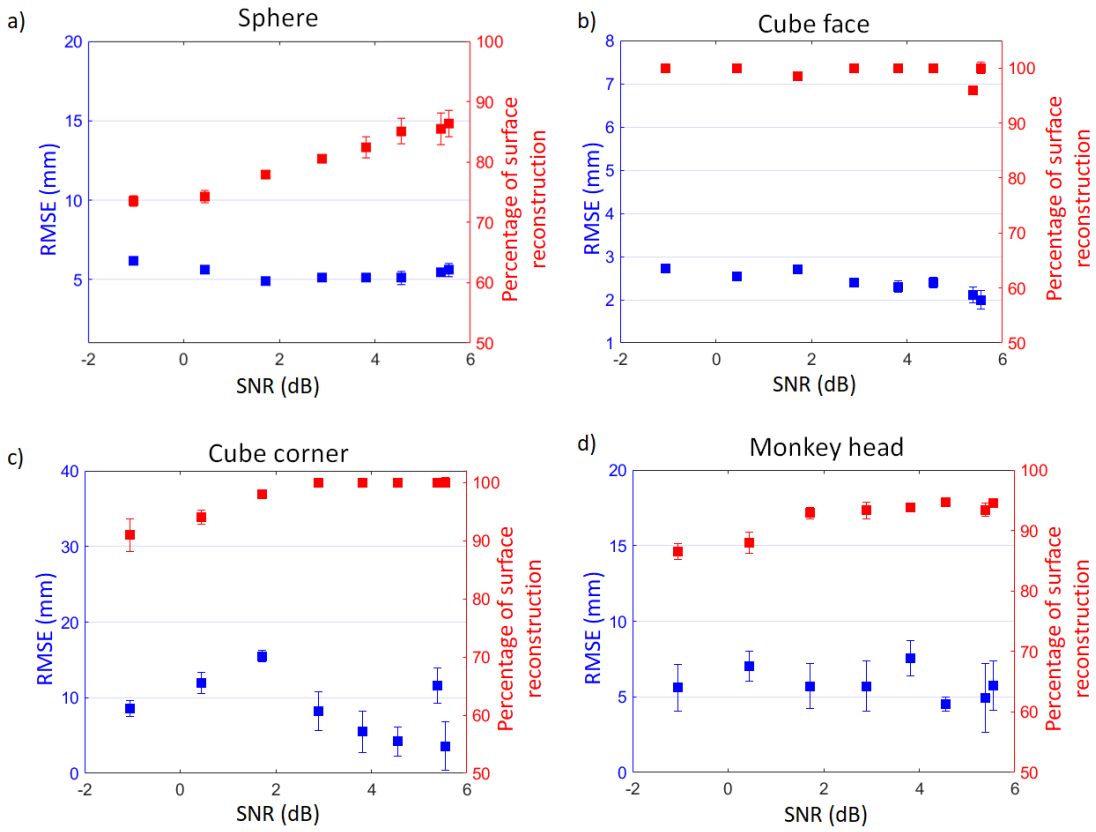


Figure 4.15: Signal to noise results. Graphs plotting the RMSE error and the percentage of the surface reconstructed regarding the signal to noise ratio for a) the sphere, b) the cube face, c) the cube corner and d) the monkey head.

the object surface. To discard the artefact values, a threshold is set in the reconstruction process where any jump, between grid points, is above a set value of 10 mm then the value at the grid point is discarded. As the light intensity decreases, the bottom area of the sphere is less illuminated which results in a loss of information in the non-illuminated area and hence high errors. To avoid high error values being incorporated in the final image, any values in the final reconstruction above the threshold are discarded. This means that as the SNR decreases, less area of the sphere can be reconstructed, but the portion that is reconstructed is not affected by the SNR. Nonetheless, 73.5 % of the object view can be reconstructed even with an SNR equals to -1 dB.

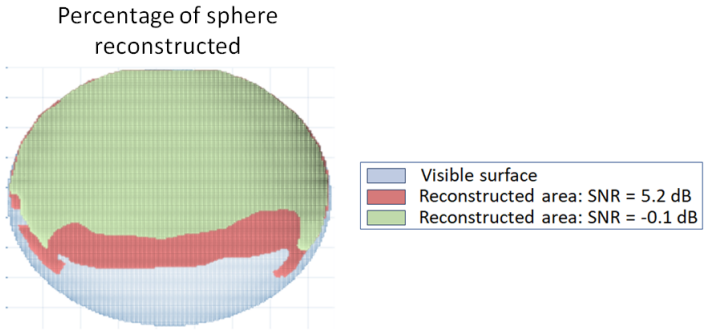


Figure 4.16: Superposition of the different reconstructions of the sphere depending on the SNR.

For the cube side, Fig. 4.15 b) shows that neither the RMSE nor the percentage of reconstruction are dependent on the SNR. The RMSE graph varies between 2 mm and 3 mm across the SNR with 100% of the cube side surface reconstructed. As the cube side is a flat surface, no shadows or lack of illumination are impacting the bottom of the cube. This means that the top-down illumination method is adapted to reconstruct flat surfaces even in negative SNR scenarios.

On the other hand, both the cube corner and the monkey head depend on the SNR for the percentage of surface reconstructed, see Fig. 4.15 c) and d). Around 2 dB, the surface percentage decreases slightly to reach 90% at -1 dB, which is considered to be a good result compared with the sphere. However, in both cases, the RMSE remains roughly constant across the SNR. The cube corner shows a variation of the RMSE between 3.6 mm to 15.5 mm which can be surprising at first. This error deviation is

mostly due to the measured gradient variation, as mentioned before, the integration time might be shorter than the frame duration and end up with different light ratio after the decoding process. Therefore, the cube angle can be more or less well reconstructed and lead to a high or small RMSE error. The cube corner graph clearly shows that this variation is independent of the SNR. Similarly, the RMSE error of the monkey head varies between 4.5 mm and 7 mm and is independent of the SNR.

These four graphs show that, depending on the shape of the imaged object, the percentage of the surface reconstructed might decrease as the SNR decreases. However, in each situation, the RMSE remains constant across the SNR which shows that the modulation scheme is robust to unmodulated background light.

4.4 Discussion

In this chapter, an accurate 2.5D reconstruction of objects with different shape complexity has been demonstrated using a new PS imaging configuration that can readily be employed in conventional room lighting scenarios. Table. 4.1 sums up the RMSE error, the NRMSE and the percentage of surface reconstructed for the sphere, the cube and the monkey head respectively. These results show that an 'out-of-plane' PS configuration can achieve similar reconstruction error to the conventional PS method [4,5,54]. Hence, for the first time, PS imaging can be applied with a camera and a lighting system set in different planes, which opens up the possibility of using the imaging technique with existing lighting infrastructure.

Table 4.1: **Error and percentage of surface reconstructed summary.**

Object	Height	RMSE	NRMSE	Surface reconstructed
Sphere	48 mm	2.69 mm	5.61%	78.4%
Cube face	75 mm	1 mm	1.36%	100%
Cube corner	75 mm	15 mm	16.6%	100%
Monkey head	80 mm	10.9 mm	18.9%	96%

Moreover, the MEB-FDMA encoding scheme allows for fast acquisition PS imaging. The modulation scheme also enables simple installation through removing the need for

synchronization, and as importantly, modulates LEDs above the visual flicker recognition threshold, thus significantly simplifying the deployment, which is not possible with successively flashed LEDs [4, 5]. Futhermore, this method can be implemented using commercially available and hand-held devices.

However, discontinuities in an object can lead to significant topography reconstruction errors and this is the main limitation in reaching high-resolution 2.5D reconstruction. Section 3.3.2 clearly shows difficulties in reconstructing the monkey head ears for example and this is a major issue for surveillance applications or robot navigation where most scenarios will contain discontinuities. A solution would be to use another 3D imaging technique that would deal with the discontinuities while using PS imaging to achieve high-resolution surface reconstruction. Chapter 6 will demonstrate such a hybrid technique where both Time-of-Flight and PS imaging will be combined.

Moreover, research on PS imaging currently focuses on achieving uncalibrated [46, 47] and non-Lambertian [48, 50–52] PS imaging in order to expand the 3D technique to real world objects and applications. As the work on synchronization-free top-down PS imaging is still at a proof-of-concept stage, the 2.5D reconstruction is carried under the Lambertian assumption. However, further work in this thesis will start to implement a deep-learning approach that will combine a faster computational reconstruction of the surface along with the removal of the Lambertian assumption. Please refer to Chapter 7 for more details.

Another 3D imaging requirement that the static PS imaging setup does not address is the real-time reconstruction of an object in motion. To fully assess the robustness of a 3D imaging method, the demonstration of a dynamic scenario is necessary. In the next chapter, the 3D reconstruction of an ellipsoid in motion will be demonstrated.

4.5 Conclusion

This chapter has demonstrated the 2.5D reconstruction of static objects using the synchronization free top-down illumination PS imaging setup with an NRMSE error ranging from 1.36% to 19% depending on the shape complexity. The error can be explained by the nature of the top-down illumination setup. The MEB-FDMA mod-

ulation scheme is easily integrated with the PS algorithm and does not require any synchronization between the mobile device and the illumination system, which simplifies its deployment. In addition, the LEDs' modulation frequency is above visual flicker recognition which makes this technique safe to use in a public environment. However, if the resolution of the 2.5D reconstruction need to be improved, one option would be to improve the quality of the four images after the decoding process. A maximum of 128 frames are currently used in the decoding stage and this is limited by the smartphone memory. With a higher memory, the time acquisition at 960 fps would increase and more frames would be recorded. The smartphone technology is always improving and more camera options will give the possibility to achieve 960 fps for longer acquisition time.

To sum up, this method can potentially be retrofitted in current building infrastructure for video surveillance applications for its simple installation, for its low-cost as it relies on commercially available and hand-held devices and for its ready-to-use aspect with a smartphone. The next chapter will focus on the demonstration of real-time 3D reconstruction of an object in motion.

Chapter 5

Dynamic Imaging

In Chapter 4, the top-down illumination PS concept in a static scenario has proven to achieve high resolution surface reconstruction with errors ranging from 5 % to 20 % depending on the complexity of the object. The goal is now to reconstruct an object in motion using the same top-down illumination PS approach.

In this chapter, the challenges of acquiring real-time reconstruction of objects in motion are first explained. Then, the adapted PS imaging setup for dynamic imaging is presented along with a full 3D reconstruction of an ellipsoid at a video rate of 25 fps. The 3D reconstruction error ranges between 4 mm and 11 mm at a distance of 42 cm and with the dimension of 60 mm, i.e between 8.8 % and 18 %, which is similar to the error obtained with the static imaging setup.

5.1 Motivation

A major motivation for the development of 3D imaging is to achieve real-time reconstruction of scenes in order to successfully reconstruct real-life scenarios such as face recognition [53, 147], machine vision [148], security and surveillance or even 'touch-less' fingerprint detection [149]. Depending on the application, real-time imaging can have different meaning. For example, robot navigation, or industrially based robot manipulation, requires multiple steps to achieve fast reconstruction that can be considered real-time. In order for a robot to sense its environment and initiate an action, the real-

time response must encompass capture of the scene, processing of the captured frames, analysis of the reconstruction and the command to send to the robot. Recently a start-up in Europe patented a new technology based on parallel structured light that improves the acquisition of dynamic scenes for dynamic robot navigation [150]. By combining the qualities of a structured light system [25, 151] and the speed of a Time-of-Flight technique [57, 72], their mosaic shutter sensor based CMOS camera can achieve a full 3D scan at 20 fps at a resolution of 2 Mega pixels [150]. Automated car self-driving is another example where real-time 3D imaging is necessary to analyse the environment for the car to drive itself, relying on different information simultaneously [30, 68].

However some other applications, such as security and surveillance, or defence, do not require the same specifications to achieve real-time 3D imaging as reconstructing the scene in 3D is the main goal. In what follows, real-time imaging is considered to encompass capture of raw frames with processing software to output the 3D reconstruction of the imaged scene. As discussed in the previous chapters, research is on-going across a range of different 3D imaging methods, where different techniques are used to reach both a capture and a computational reconstruction time that can be considered real-time. The same challenge applies to all 3D imaging methods; the higher the camera raw frame rate, the faster the effective reconstruction frame rate. This means that a trade-off must be found between the exposure time and the noise as a high frame rate results in low integration time, and low integration time results in low-photon flux detection which causes low Signal-to-Noise ratio (SNR) and hence affects the accuracy of the reconstruction [27, 152]. Therefore, conditions, in term of the acquisition speed and exposure time, apply on sensors to complete dynamic imaging.

LiDAR technology relies on high-sensitivity of a photon counting detector along with fast timing electronics [27], which makes it a good option for real-time imaging. As shown in Table 5.1, this technique can achieve 3 fps real-time 3D reconstruction, which can be fast enough for video surveillance applications [27]. However some mathematical processing of LiDAR data can improve this reconstruction rate up to 50 fps [153]. This recently developed computational framework, which takes advantage of the LiDAR database point clouds, can achieve reliable real-time 3D reconstruction from single-

photon LiDAR at a rate of 50 fps using a point cloud denoiser [153]. Some additional or complementary methods, beside using a high-speed sensor, can be needed to reach surface reconstruction within video frame rate. For example, a 3D scanning system combines stereovision and active illumination based on phase-shift to capture accurate depth maps of complex dynamic scenes at 17 fps [154]. However, some additional synchronization processes are then needed to calibrate multiple cameras together, which is something that should be avoided when making an easily deployable imaging system.

Table 5.1: **Real-time imaging comparison**

	3D imaging Method	Acquisition rate (camera raw frame rate)	Effective frame rate	Illumination modulation rate	Processing delay per frame
Tachella *et.al* [153]	LiDAR	150,000 fps	50 fps	150 kHz pulsed laser	20 ms
Edgar *et.al* [27]	LiDAR	20 kHz histogram	3 fps	100 MHz pulsed laser	
Herrnsdorf *et.al* [54]	PS	60 fps	15 fps	60 Hz	
Schindler *et.al* [53]	PS	10 fps	10 fps		
Yoda *et.al* [152]	PS	75 Hz per 3 images	70.5 fps	75 Hz	
Malzbender *et.al* [155]	PS	500 fps	up to 60 fps	62 Hz	0.14 s
Hernandez *et.al* [139]	PS with colored light	60 fps	20 fps		20 s
Vlasic *et.al* [3]	Multi-view PS	240 fps	60 fps	60 Hz	
Weise *et.al* [154]	Stereo and active illumination	120 fps	17 fps	120 Hz	
Our work	**PS**	**1,000 fps**	**25 fps**	**1,000 Hz**	**5 s**

It is therefore clear that a trade-off must be found between the use of a high speed sensor along with an easily deployable 3D imaging technique. Applying dynamic imaging to PS can be a challenge because of synchronization issues between the light position according to the scene as relative movement must be controlled [156], but also because the scene must remain static to acquire at least the three images needed for the estimation of the surface normal [152]. Nonetheless, some previous work shows that it is possible to adapt PS imaging to reconstruct the shape of object in motion. For example, a coloured-light approach shows that multi-spectral PS can recover a depth map at 60 fps [139]. Other techniques make use of high-speed cameras to obtain a fast effective frame rate such as [3, 155] where they respectively use a 240 fps and 500 fps camera to both achieve a surface reconstruction with an effective frame rate of 60 fps. Complementarily, the development of new sensors along with the release of high-performance

GPUs helps in reaching faster acquisition and data processing leading to faster effective frame rate. As shown in Table 5.1, current effective frame rates range between 3 fps to a maximum of 70.5 fps with a processing delay varying between 20 ms to 20 s. In [152] is presented a dynamic PS method that uses a new sensor called a 'multi-tap CMOS image sensor' that divides the electrons from the photodiode from a single pixel into different bins and thus captures multi-images under different lighting conditions with almost identical timing. Acquiring multi-images at the same time improves the acquisition speed and therefore improves the effective frame rate of the determination of the surface normal. However, the sensor in [152] needs to be synchronized with the light sources and does not recover the full shape reconstruction of the scene in real-time.

This chapter will demonstrate that the light modulation scheme and PS method [142] can be adapted and applied to dynamic imaging while keeping the synchronization free aspect. Some trade-offs need to be considered. First the acquisition speed of the camera must be high enough to result in an effective frame rate within 20 fps range. When the camera frame rate is high, the light level from the LEDs must be powerful enough to capture enough photons during the integration time. In addition, the speed or motion of the imaged object must be adapted to the acquisition frame, hence integration time, to avoid any motion blur. An important consequence is therefore on the choice of camera, which must achieve a high acquisition rate with a memory big enough to store the raw data. Unfortunately, the smartphone used in Chapter 4 limits the acquisition time to 0.2 s because of its memory storage. A high performance camera is therefore used in what follows as its acquisition frame rate and memory are high and adapted to dynamic imaging. High-speed cameras are available but mostly used as laboratory equipment because of their high cost. However, the continuing development of smartphones will make mobile phone devices suitable and portable for dynamic imaging in the near future. With this high-speed camera, a camera raw frame rate of 1 kfps is reported with an effective frame rate of 25 fps for an error ranging between 6 % and 18 %. As the 3D reconstruction process is done off line at a rate of 3 min, per effective frame rate, which is equal to 5 s per frame, the dynamic imaging approach cannot yet be considered real-time. Nonetheless, the acquisition speed allows us to achieve

a fast enough effective frame which will allow us, in future work, to reach real-time reconstruction after speeding up the 3D reconstruction process.

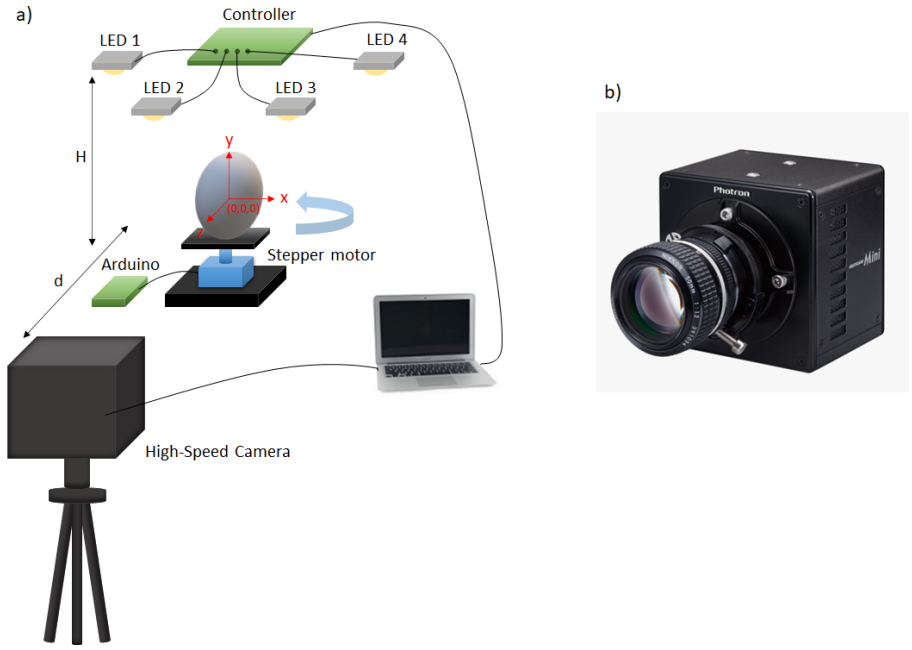


Figure 5.1: a) Schematic of the experimental setup in the 'out-of-plane' configuration for dynamic imaging, b) Picture of the Photron MiniUx100 high-speed camera used in the setup.

5.2 Dynamic imaging setup

The top-down illumination setup detailed in Chapter 4 for static imaging is modified slightly here to meet the requirements for dynamic imaging. As shown in Fig.5.1 a), a stepper motor (RS PRO Hybrid, Permanent Magnet Stepper Motor, 1.8° step) is added in order to rotate a 3D printed ellipsoid which is 100 mm long and 60 mm wide. A high-speed camera (Photron MiniUx100, see Fig.5.1 b), replaces the mobile phone due to its larger video capture memory thus enabling a longer acquisition time. The camera frame rate is set at 1 kfps with a shutter speed of 1 ms and is matched to the LED modulation rate. As the acquisition rate is not modified, the LED illumination level remains the same as in Chapter 4, i.e. a measured optical power of 20 μW at the object location. The stepper motor rotates at a speed of 7.5 RPM, which is the calculated

rotation speed to capture enough frames per ellipsoid position while avoiding motion blur. The acquisition runs for 8 s and the real-time video of the ellipsoid in motion can be found under the folder 'Dynamic imaging' in [157]. The 3D reconstruction is done off line with the same reconstruction program pipeline as demonstrated in Chapter 4, which is shown again in Fig. 5.2. The reconstruction requires at least 32 frames for a full reconstruction and here 40 frames are recorded for each 3D frame to match the motor step duration and to achieve effective full 3D reconstruction at a video rate of 25 fps.

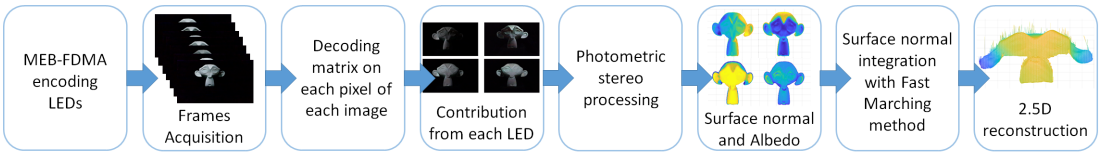


Figure 5.2: Acquisition and Reconstruction program pipeline. LEDs are encoded with an MEB-FDMA scheme; the mobile device acquires a stack of images that are demodulated with a decoding matrix; four output images are then retrieved: one for each illumination direction; the photometric stereo processing determines the surface normal components and the albedo; afterwards the surface normals are integrated with a Fast Marching method to then obtain the 2.5D reconstruction of the object.

By running the raw camera acquisition at 1 kfps for 8 s, a total of 8,000 raw images need to be processed to reconstruct the full rotation of the ellipsoid, which consists of 200 different single positions. As pictured in Fig. 5.3, 40 frames are used per single ellipsoid position which means that the achievable effective frame is equal to 25 fps, when running the camera acquisition at 1 kfps. The effective frame rate can be increased when increasing the acquisition rate, although a balance needs to be found between the acquisition speed, the illumination level and the rotation speed of the object. The current reconstruction pipeline takes about 3 min to determine the 3D reconstruction per ellipsoid position, which makes it time consuming to process. In order to speed up the reconstruction time of the full ellipsoid rotation and to allow error calculation, the full 3D reconstruction analysis is completed on 21 ellipsoid topographies out of 200 available. Each of the 21 reconstructions are split by an angle of 18° in the rotation of the object. In order to have a smooth display, the surface normal components and the 3D reconstructions are displayed at a frame rate of 3 fps. This video display rate

is 10 times slower than the effective rate that can be achieved but it still represents a 3D reconstruction video matching the object in motion. Two videos, one for the surface normal components and one for 3D reconstruction, can be found in the folder 'Dynamic imaging' in [157]. These two videos cannot be considered real-time as the reconstruction pipeline needs to be executed faster. However, the effective frame rate of 25 fps demonstrates the potential to reach a video rate real-time display of 3D reconstructed scenes as soon as the computational time is improved.

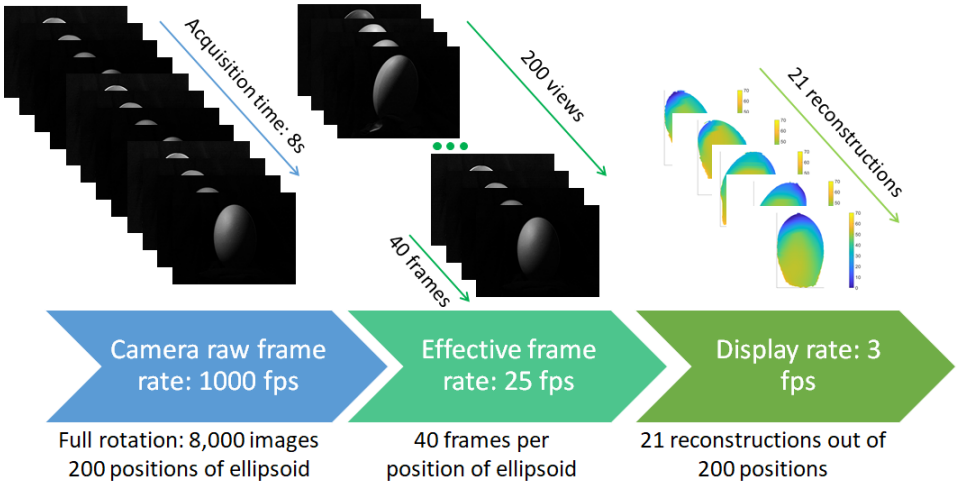


Figure 5.3: Flowchart explaining the determination of the effective frame rate. 8,000 images are captured at a camera raw frame rate of 1 kfps in which 40 images are used, per ellipsoid position, for the 3D reconstruction. To allow error calculation and decrease computational work, 21 ellipsoid positions are 3D reconstructed and displayed at 3 fps.

5.3 Results

Fig.5.4 shows the extracted images for the different LED illumination directions, for five different views of the ellipsoid in motion. Similar to the procedure described in Chapter 4 and shown in Fig. 5.2, a decoding matrix is applied on 32 images from the 3D video frames. However, as explained in section 4.2, 40 images are available per 3D video frame, which means that 32 images need to be selected out of the 40 images available, please see Fig. 5.5. This strategy has been adopted to avoid synchronising the camera with the stepper motor. Therefore, any possible overlapping of images between two

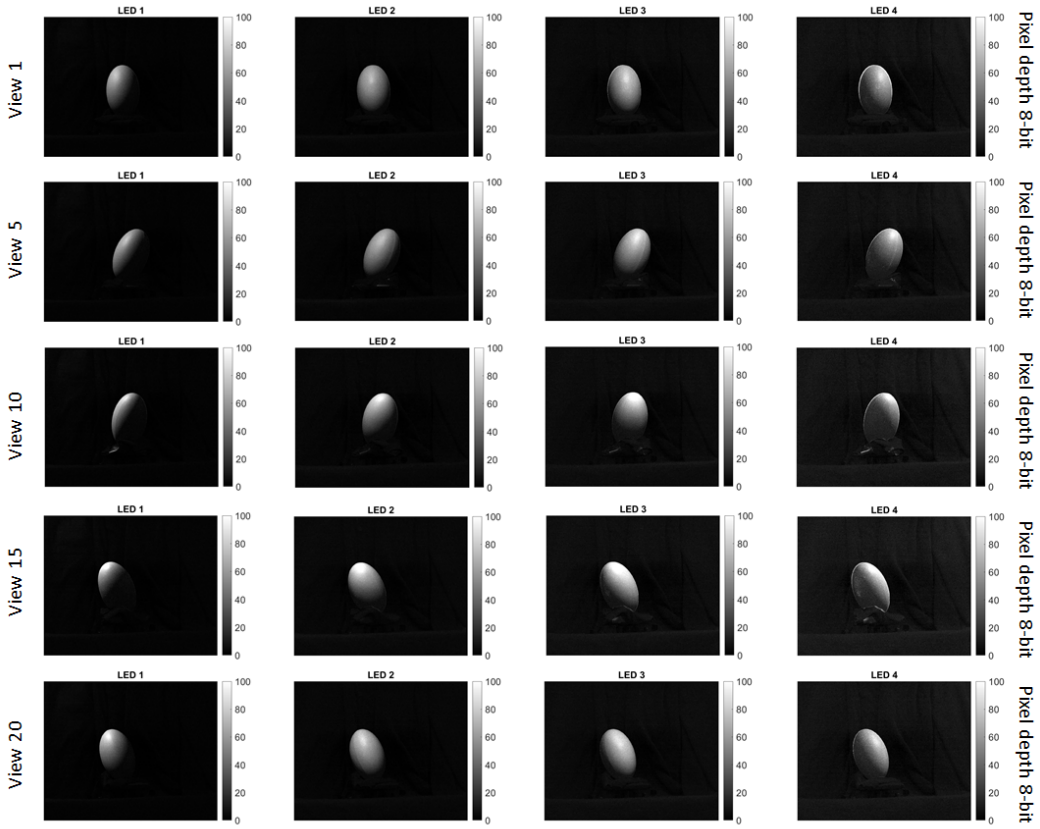


Figure 5.4: Decoded images of the ellipsoid for 5 different views. Obtained after demodulation of the recorded frames for LED1, LED2, LED3 and LED4.

consecutive ellipsoid positions needs to be minimized. By discarding the first and last 4 images to end up with the 32 images required, any synchronisation error is kept to a minimum. Indeed, any small error in the synchronisation between the captured frames and the stepper motor will not have a big impact on the decoded images. By looking at the extracted images in Fig.5.4, no motion blur can be visualised. It is important to notice that the ellipsoid is constantly moving, and despite its motion, its boundaries are sharp and well-defined which shows that the imaging rate is adequate. In addition, Fig.5.4 shows the different illumination directions of the four LEDs for different views.

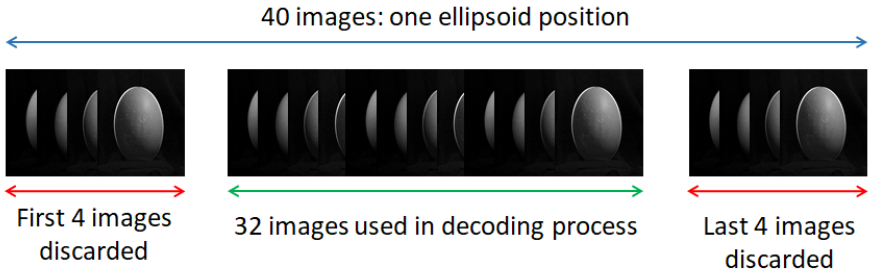


Figure 5.5: Schematic explaining the selection of images to run the decoding matrix for dynamic imaging. 40 images are available per 3D video frames and only 32 images are used during the decoded process, hence the first and last 4 images of the sequence are discarded. This strategy removes the need to synchronise the object in motion with the camera.

The surface normals behave very similarly to the static situation in that a high fidelity is observed in N_x, while N_y and N_z have lower but still useful fidelity, as shown in Fig.5.6. The albedo shows some higher values in the top area of the ellipsoid and can be explained by, first, the top illumination aspect of the setup and, second, the material of the ellipsoid. Comparing to the objects used in Chapter 4, the material of the ellipsoid is slightly shinier and shows some specular highlights as observed in Fig.5.4 also. However, as the specular effect does not impact the surface normal components, the Lambertian assumption is kept when applying the photometric stereo algorithm.

The 3D reconstruction video shows the 21 reconstructed frames with a colour-coded plot and a rendered view, which are repeated three times, and displayed at 3 fps. The video is available in the folder 'Dynamic Imaging' in [157]. Snapshots from the video are displayed in Fig. 5.7 for five different views, where the surface reconstruction, the

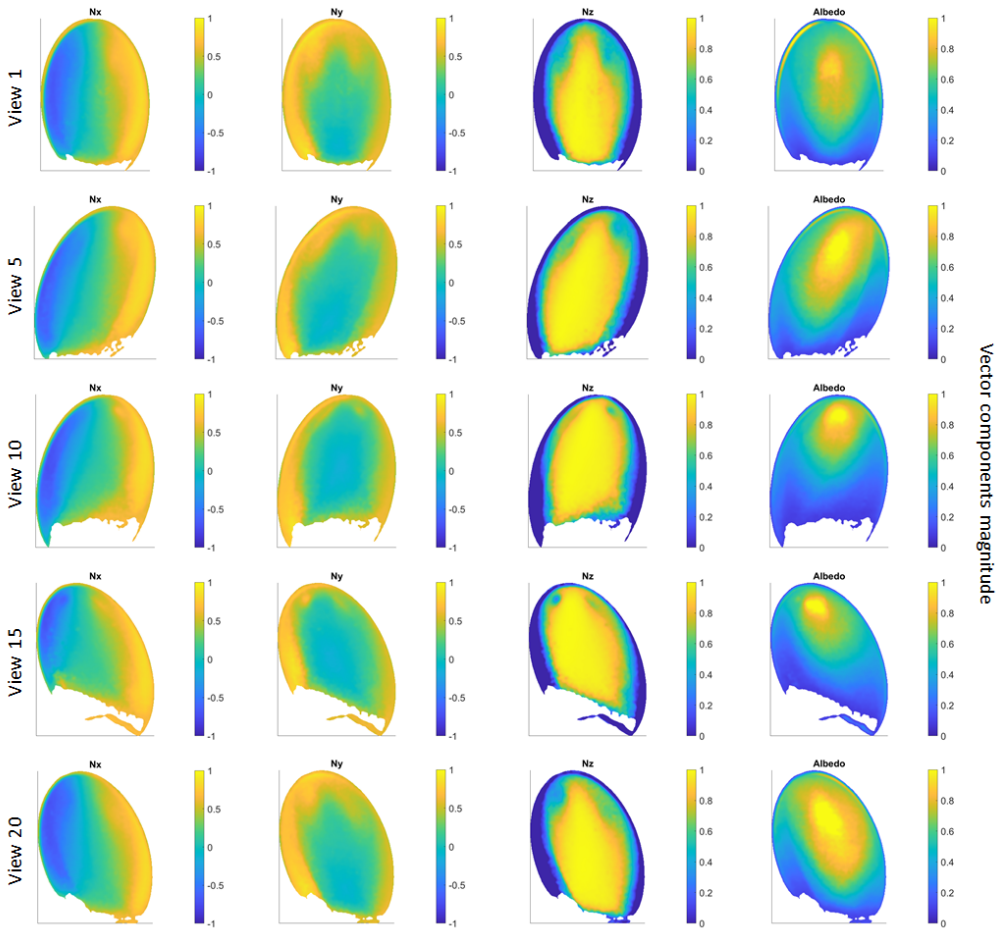


Figure 5.6: Surface normal components and albedo of the ellipsoid for 5 different views. Obtained after running the photometric stereo algorithm, respectively Nx, Ny, Nz and Albedo.

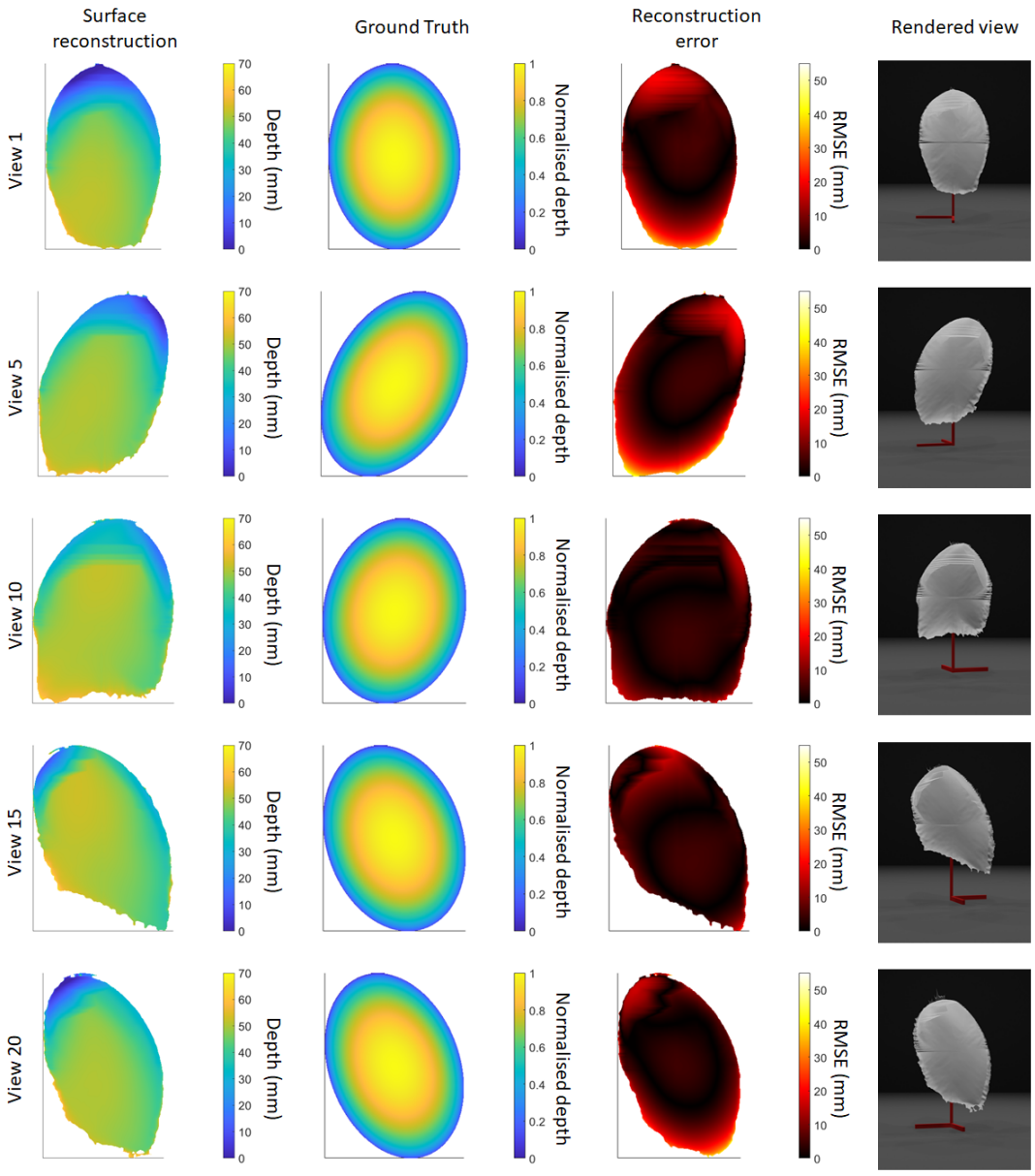


Figure 5.7: Surface reconstruction of the ellipsoid for 5 different views. The first column corresponds to the 3D reconstruction of the ellipsoid, the second column plots the normalised ground truth, the third column shows the RMSE error map and the last column is the rendered view obtained from Blender^{TM}.

ground truth, the RMSE error and the rendered view are plotted. Globally, the gradient observed on the 3D reconstruction matches the one from the ground truth, as the depth of the reconstructed surface decreases from the centre to the edges. The ground truth was normalised and adapted to the 3D reconstructed shape values to easily determine the RMSE and NRMSE error, please see Chapter 4 for more details.

Table 5.2: **RMSE, NRMSE, Ratio Mask and Ratio Ground Truth of the ellipsoid for five different views**

	View 1	View 5	View 10	View 15	View 20
RMSE	11 mm	11 mm	4.8 mm	5.58 mm	9 mm
NRMSE	18.5%	20%	12%	12.5%	15.6%
Ratio Mask	80.7%	83.3%	79.4%	82.1%	83.4%
Ratio Ground Truth	81.7%	86.5%	72.6%	69.1%	81.1%

The RMSE error map has been calculated for the five different views and the map shows that the error is similar across the different views, see Fig. 5.7. Indeed, the RMSE error in the centre of the ellipsoid ranges between 0 and 10 mm error while the error on the edges of the ellipsoid can increase up to 40 mm. The artefacts on the edges can also be seen on the rendered view, see Fig. 5.7 in the right hand column, and are mostly caused by poor numerical conditions due to $N_z \approx 0$. A flat reconstruction at the bottom of the object is also visible which is expected with the top-down illumination. Similarly to the sphere result with the static configuration, the ellipsoid is not entirely reconstructed. Some of the bottom part is missing which is not only due to the top-down illumination but also to the support piece that has been used to hold the ellipsoid in a tilted position. Nonetheless, the global 2.5D reconstruction of each view is satisfactory and of comparable quality to the static scenes. Table. 5.2 sums up the calculated RMSE and NRMSE error for each displayed view and the results vary between the 3 mm/11 mm range obtained in the previous chapter. Table. 5.2 also shows that some views have lower RMSE error than others which can be explained by the percentage of the area reconstructed, shown by the ratio mask and ratio ground truth. Ratio Mask is the proportion of surface reconstructed across the mask that is below a defined threshold, *i.e.* around 3 mm above height of the object, so that artefacts are not considered in the error calculation. Ratio Ground Truth is the proportion of surface

reconstructed across the entire ground truth of the object and that is also below the threshold. The proportion of ellipsoid reconstructed on View 10 and View 15 is smaller than the other views as the support piece hides a part of the object. Therefore, the error obtained at the bottom of ellipsoid are not accounted for in that case, which reduces the global error.

5.4 Discussion

Our MEB-FDMA modulation scheme and top-down illumination PS imaging setup can thus easily be applied to dynamic imaging without having to bring many modifications to the experimental setup and to the reconstruction pipeline. The reconstruction error obtained ranges between 4 mm and 11 mm at a distance of 42 cm and with dimensions of 60 mm, *i.e.* 6.6 % and 18 %, which is similar to the static imaging results. Furthermore, no synchronisation between the camera and the stepper motor is needed and 3D reconstruction results show no motion blur or major artefacts, which means that the light level and the rotation speed of the object is adapted. However, frames were discarded to avoid having to synchronise the camera with the stepper motor. A solution would be to use frame differencing to track any change in position of the object. With frame differencing, a new frame would always be compared to the previous one. Once the object moves to a new position then the frame selection would be triggered for the surface reconstruction of the new position.

The high-speed camera acquisition is also fast enough to reach an effective frame rate of 25 fps, which aligns with most of the results obtained in the literature except for [3,152] which can reach twice the rate or more. It is clear that the absolute effective frame rate is derived from the camera acquisition rate which makes it dependent on the hardware performance. With today's evolution in smartphone and camera technologies, the effective frame rate is likely to improve in the future to reach faster real-time performance, which will make the technology interesting for such applications as indoor robot navigation for public services.

However, the computational reconstruction time is the main limitation in reaching real-time dynamic imaging. The current reconstruction pipeline is time-consuming for

accessible imaging processing electronic systems, which does not allow for the reconstructed 3D shape to be displayed at the effective frame rate. Nonetheless, if a faster algorithm were to be used, the technique demonstrated in this chapter can achieve an effective frame rate of 25 fps using a synchronisation free version of PS imaging that, while lower in absolute frame rate than previous demonstrations [142], is easier to implement with discrete hardware components and is robust to background illumination noise, as detailed in Chapter 4.

To improve the computational time, a neural network algorithm could be used which would give the opportunity to input a video captured with a mobile device and to directly output the 3D shape of an object or a scene. Current research on deep-Photometric Stereo [18, 51] shows encouraging results in term of reconstruction speed. Using a neural network would also show potential in improving the PS setup by achieving non-calibrated PS imaging of non-Lambertian surfaces [17, 51]. Some premises of the deep-learning work can be found in Chapter 7.

5.5 Conclusion

This chapter has demonstrated the surface reconstruction of an ellipsoid in motion under the top-down illumination PS imaging setup using the MEB-FDMA modulation scheme. Results show sharp boundaries and well-defined edges shape reconstruction with an error ranging from 8.8 % to 18 % which goes along with NRMSE results from static imaging. The error can be explained by the nature of the top-down illumination setup.

No synchronisation between the object in motion and the camera was required. However, the current approach of frame selection may not be the most robust and can become an issue if objects were to move faster than the ideal speed measured. Some techniques such as structure-from-motion or frames differencing would be useful in determining the movement of the object and from it select the frames to be decoded for each position. Indeed, in structure-from-motion, both spatial and temporal changes are monitored which would mean than more information from the motion of the object can be derived. A higher quality image would then be derived as less error in the frame

selection would impact the determination of the surface normal vector.

Nonetheless, with the current approach, the effective frame rate achieved is 25 fps video rate. Because of the computational time requirements of the current reconstruction pipeline real-time imaging is not yet achievable. Therefore, further work will focus on improving the computational speed of the current method by using a deep-learning approach.

Chapter 6

Hybrid imaging

Although a high-resolution surface reconstruction was achieved in Chapter 4 using different shape complexity objects, discontinuity issues made it difficult to obtain high-fidelity results in those areas. To deal with the discontinuities, two 3D imaging methods with complementary properties, Time of flight and PS, can be combined. The former can achieve depth accuracy in discontinuous scenes and the latter can reconstruct surfaces of objects with fine depth details and high spatial resolution. This chapter demonstrates the surface reconstruction of complex 3D fields with discontinuity between objects by combining the two imaging methods. Using commercial LEDs, a single-photon avalanche diode camera, and a mobile phone device, high resolution of surface reconstruction is achieved with an RMS error ranging between 4% and 5% for an object auto-selected from a scene imaged at a distance of 50 cm to 70 cm. The results presented in this chapter have been published: E. Le Francois *et al.*, Optic Letters, 46, 3612 (2021).

6.1 Motivation

3D imaging is becoming well-established using a range of different techniques, namely time-of-flight (ToF) [23, 56, 57, 62], light detection and ranging (LiDAR) [27, 69] and photometric stereo imaging (PS) [4,39,158], please refer to Chapter 1 for more details. A trade-off is usually needed between some 3D imaging requirements such as the depth

accuracy, the spatial resolution, the capture speed or the processing time. None of the 3D imaging techniques encompass all the requirements at once, which can be an issue when a high spatial resolution is required to image a scene with discontinuities. For high fidelity results, several techniques may be combined, such as ToF and PS, which complement each other particularly well.

ToF relies on the time-correlation of a reflected optical signal with the outgoing pulse to obtain a range map of the illuminated scene [57]. Laser-based ToF excels at long range target recognition, because the signal-to-noise ratio limited range can be as much as 45 km [69] and the depth resolution is determined by the timing accuracies of the output pulse and the detector. Laser-based techniques can achieve high timing accuracies allowing sub-cm accuracy at km ranges [23], while light-emitting diode (LED) sources [62] achieves cm-scale depth resolution. Recent advances in ToF were enabled by single-photon avalanche diode (SPAD) detectors fabricated in a complementary metal oxide semiconductor (CMOS) process with on-chip digital logic [56, 159, 160]. Each pixel of a SPAD array can extract a distance measurement from a light pulse emitted on a scene and reflected back on the sensor [56, 62]. LEDs are becoming widely used in 3D imaging for their achievable high-modulation rate, compact size, eye-safety and power-efficiency. LED-based ToF imaging can achieve high-depth accuracy of discontinuous scenes but is limited by the resolution of current single photon camera systems, or acquisition time of scanning systems [62].

On the other hand, PS [39] uses conventional imagers and therefore has a high spatial resolution in all three dimensions [158]. As demonstrated in Chapter 4, this technique does not deal well with discontinuities nor give absolute range, but PS can easily be applied to dynamic scenes and achieve real-time imaging [142] (see Chapter 5). It is therefore clear that ToF and PS are complementary 3D imaging techniques that can potentially achieve high-fidelity results by merging the two.

Sensor fusion has been investigated in the past few years mainly to improve the low-resolution depth of ToF cameras by using high-resolution intensity images [73]. Sensor fusion can also be used to improve 3D mapping procedures with RGB-D cameras [74,75] or perform segmentation and tracking [76]. Because of the impact of random noise [63],

fine depth details are lost when using a ToF camera. However, PS imaging is robust to noise and can provide finer depth details than ToF [9] but PS does not provide absolute distances. By fusing such approaches together, the technique takes advantage of the spatial redundancy. Another approach would be to use an image super-resolution technique to utilize temporal redundancy [77]. To date, most of current sensor fusion techniques rely on improving the ToF sensor depth map by measuring both a range map and a surface normal map to then merge it to improve the resolution [9, 38, 77]. However, in this chapter, the main focus is to improve the PS imaging technique by using a range map as a masking tool to deal with discontinuities. The advantage of this approach is that a fusion scheme does not need to be developed and only a calibration between the two sensors would be needed to then use the exact same PS imaging process used in Chapter 4. By employing a dual imaging system incorporating both ToF and PS, the complementary properties of both systems can be used to image complex 3D fields with high-resolution and complex discontinuities between objects.

This chapter reports the 3D reconstruction of different objects placed at different depths within the image scene using a combination of ToF and PS imaging, where ToF enables detection of the boundaries and absolute depth of each object while PS imaging yields a high resolution surface profile within these boundaries. A time-correlated single photon-counting (TCSPC) SPAD camera [108] is used with pulsed blue LEDs to obtain a range map of the scene. For PS imaging, four modulated white LEDs illuminate the object while a mobile phone captures frames at 960 fps. The LEDs are modulated, with the same modulation scheme as Chapter 4, at the camera frame rate. Again, no electronic synchronization is needed between the LEDs and the phone [142] and the visible flicker is at a minimum and not visible to the human eye. The results of the surface reconstruction show a root mean square error (RMSE) ranging from 4% to 5%.

6.2 SPAD configuration

The SPAD image sensor used here was designed by Robert Henderson's group at the University of Edinburgh and the firmware was developed by David Li's group at the University of Strathclyde. The SPAD consists of 192x128 pixels implemented in 40-nm

CMOS technology with dedicated timing electronics [108]. Each pixel is 18.4 x 9.2 μm^2 in area and can be operated with the Time-Correlated Single Photon Counting (TCSPC) functionality. Each pixel has a 33 ps resolution, enabled by a time-to-digital converter (TDC) coupled to the SPAD, and can therefore record the time-of-arrival for the first detected photon in an exposure period [108]. Each output frame consists of a time bin value for each pixel. Repeating this over many frames allows a histogram of arrival times to be built up in a pixel-wise manner, allowing 2-D imaged TCSPC to be performed [62]. The SPAD sensor can run under two modes: Photon Counting and TCSPC, where both modes are addressed and read through a Field Programmable Gate Array (FPGA) module (Opal Kelly XEM6310-LX150) with user control provided through a graphical user interface. A picture of the sensor mounted on a PCB is shown in Fig. 6.1 with the FPGA shown on the back of the SPAD.

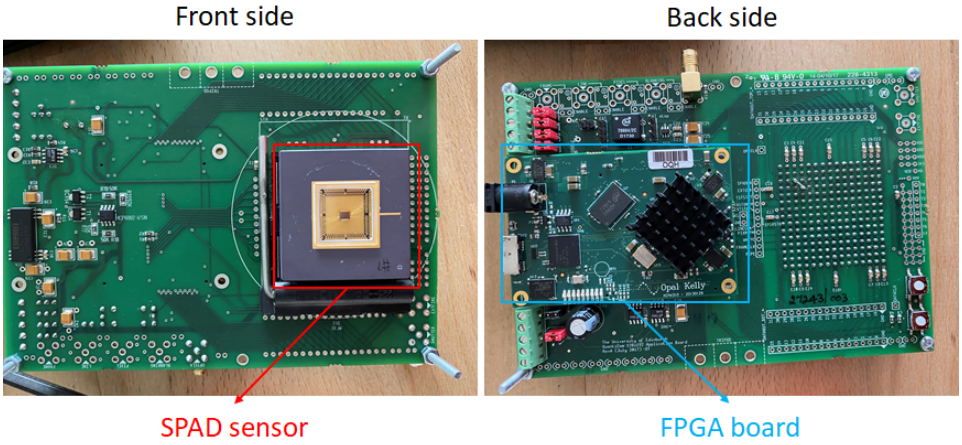


Figure 6.1: Picture of the SPAD camera mounted on a PCB motherboard. The SPAD chip 3.15 mm x 2.37 mm in size. An FPGA board controls the SPAD camera.

6.2.1 Photon Counting mode

The Photon Counting mode of the SPAD camera is used to acquire an intensity image of a chequerboard pattern for the calibration stage, which is explained in section 5.4. A total of 200 frames are acquired for one intensity image with an exposure time set at 10 ms, so the frames are therefore acquired at 100 fps. The acquisition rate does not need to be high-speed as the calibration is run one time only per object location and

then the calibration parameters are stored to then be applied to the depth-mask. By choosing a large enough exposure time, more photons can be detected which makes for a better quality intensity image, hence a better detection of the chequerboard in the calibration process.

6.2.2 Time-correlated single-photon detection mode

The TCSPC mode of the SPAD camera is used to acquire the scene depth map which will then be used as a masking tool during the PS stage. To avoid any Poissonian noise, which dominates in low photon-count images, 10,000 TCSPC frames are acquired at 480 fps for 21 s. The acquistion rate is limited by the current implementation of the FPGA, which is not optimised to use the maximum frame rate achievable with the chip (18.6 kfps [108]). Each frame is from an exposure of 1 ms, which covers many repetitions of the waveform as the trigger signal is sent every 320 ns. The SPAD sensor operates in reverse mode TCSPC, meaning the detected photon starts the timing circuitry, and an on-chip clock provides a stop signal. Therefore, the SPAD camera records timing information for only the first photon to arrive within the exposure [62, 108], producing single time-of-arrival value per frame for each pixel.

Due to the time taken for photons to travel from the LEDs to the scene and back to the camera, the entire waveform will be shifted in time depending on the distance to the imaged objects. The distance between each object in this exemplar system is about 10 cm which corresponds to a temporal shift below 1 ns. It is a challenge to measure this very narrow pulse shift and to achieve it a cross-correlation approach has been developed earlier by the group at the Institute of Photonics to improve precision [62]. By taking a reference histogram at a known distance, the time shift can be determined accurately by cross-correlating it with the scene histogram. More details on the method can be found in [62]. During the TCSPC measurement of the scene, a white reference board is therefore placed at 60 cm from the SPAD camera to measure the reference histogram. The white board is then removed to permit measurement of the scene. The TCSPC histogram is constructed and analysed off-line on a laptop. By using this method, sub-ns scale resolution can be achieved, resulting in 3.4 cm depth resolution.

By performing this process on a pixel-by-pixel basis, a depth map can be determined across the scene.

6.3 Photometric stereo

The surface reconstruction pipeline of the PS imaging remains the same as in Chapter 4. Four white LEDs illuminate the scene from four distinct illumination directions and a mobile phone records the frames at a super slow-motion mode. The LEDs are electronically interfaced with a distinct modulation fingerprint for each, so the camera can readily determine which LED generated which image, thus facilitating the 3D reconstruction. The modulation scheme is self-clocking so the camera operation can be self-synchronized with the LEDs. The recorded stack of frames are demodulated off-line on a laptop and generate four distinct images, one for each LED. The obtained images are processed by a PS algorithm to determine the surface normal components and the albedo. The surface normal components are then integrated only within the corresponding mask, which speeds up the surface reconstruction [116, 142].

6.4 Calibration method

The SPAD camera and the mobile phone have different resolution and different pixel size. In this experiment, both sensors are calibrated via a one-time calibration of the image depth using a screen with a chequerboard pattern at each object location. Fig. 6.2 displays the intensity images of the chequerboard obtained with the SPAD and the mobile phone, respectively. In order to create a bespoke mask for the PS process, the SPAD depth-map must be scaled to the mobile phone images. The chequerboard patterns can be identified automatically on MATLAB^{TM} to calculate the relative pixel scales. Once the SPAD is scaled, the chequerboard patterns from the new scaled SPAD chequerboard are detected to then proceed to the frame alignment of both sensors. Fig. 6.3 shows the superposition of the smartphone chequerboard in purple with the scaled SPAD chequerboard in green. Parameters from this calibration are saved and applied to the mask obtained in the imaging process described in section 5.6.1. In real

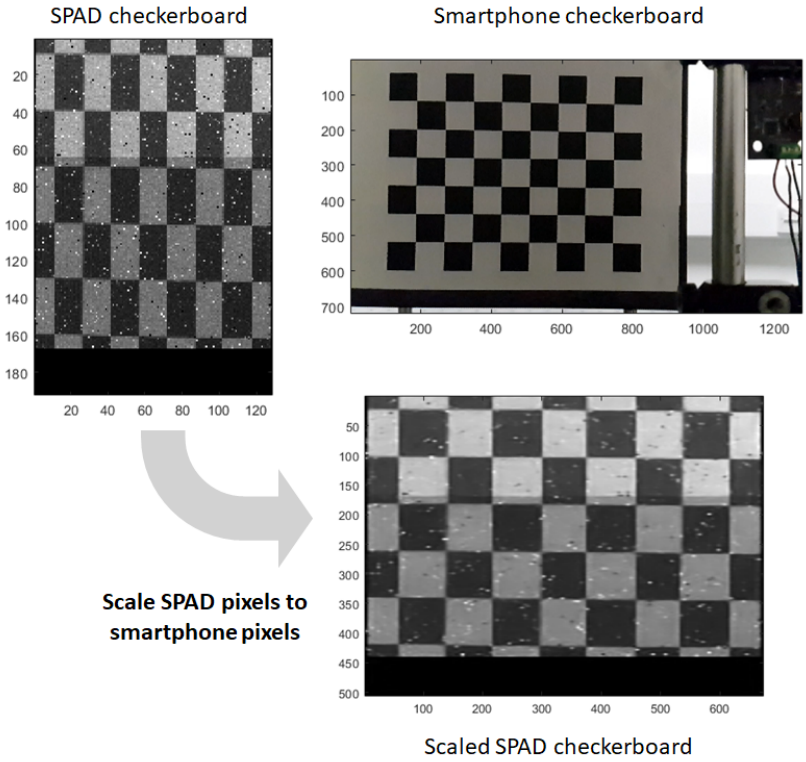


Figure 6.2: Scaling step in the calibration process. The SPAD chequerboard, on top left, is used to scale the SPAD pixels to the mobile phone pixels with the smartphone chequerboard, on top right. Chequerboard points are detected to calculate the relative pixel scales which is applied to the SPAD image to obtain the scaled SPAD chequerboard on the bottom right.

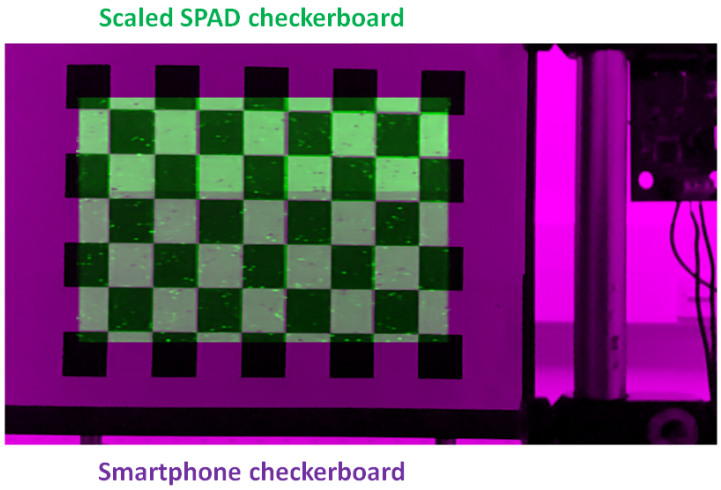


Figure 6.3: Superposition of the scaled SPAD chequerboard with the smartphone chequerboard after frame alignment.

life application, the ToF range map will be used to select the correct calibration from a pre-determined calibration table.

6.5 Optical acquisition system

As illustrated in Fig. 6.4, the ToF system comprises a 192x128 pixel SPAD sensor placed at 80 cm from the sphere, of which 167x128 pixels are used as the bottom rows pixels died and were not usable. The SPAD camera is fitted with a set of optics with a focal length of 8 mm (Navitar MVL8M23), providing a Field-of-View (FOV) of 16 degrees. Upon a trigger signal from the SPAD camera, the blue LED array (Lumileds Luxeon Z series) emits a train of 10 pulses where the pulses are driven with an electrical waveform of 11.3 ns pulse width and a pulse spacing of 30 ns. Both LED array and SPAD camera are controlled with FPGA modules (Opal Kelly XEM6310-LX150).

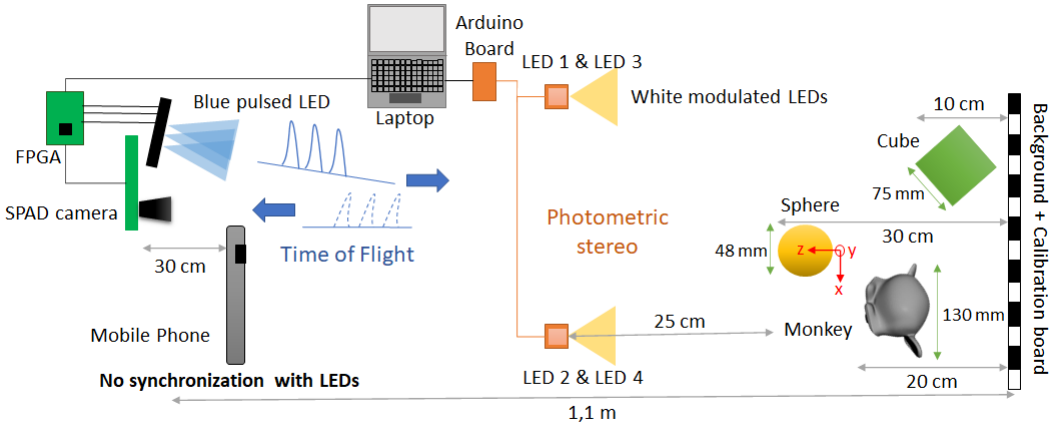


Figure 6.4: Schematic of the experimental setup.

The PS experimental setup consists of a mobile phone device (Samsung Galaxy 9) mounted on a tripod at a distance of 50 cm with a FOV of 32 degrees. Four commercial white LEDs (Osram OSTAR Stage LE RTUDW S2W) were placed 25 cm away from the object in an X shape and were connected to a controller board (Arduino Uno) for the LED modulation. A series of geometric solids were 3D printed, namely a sphere with a 48 mm diameter, a cube which is 75 mm wide, and a complex shape of a monkey head that is 130 x 94.5 mm^2 wide and 79 mm deep (see Chapter 4). On the setup, the

geometric centre of the scene is the reference (0,0,0) and the location of the LEDs is determined from this reference point. The LEDs are located at (x,y,z): LED1 (-14.5, 9, 25), LED2 (-14.5, -5, 25), LED3 (14.5, 9, 25) and LED4 (14.5, -5, 25), all in cm. Each LED was modulated with the MEB-FDMA signal (see Chapter 3) at a on-off keying rate of 960 b/s, which is above visual flicker recognition. The carriers were designed such that, analogous to orthogonal frequency division multiple access, no synchronization between the LEDs and the mobile phone was required [142], (see Chapter 3). The phone captured frames at a rate of 960 fps for 0.2 s where the acquisition time is determined by the phone memory.

Fig. 6.5 shows pictures of the experimental setup, with on the left a 'working from home' situation where the experiment was first set up. On the right is a picture of the setup taken in the lab that shows the smartphone and the SPAD camera belonging in different plane. The calibration stage gives us the possibility to set the two sensors at different distances from the scene. To achieve high-resolution reconstruction, the mobile phone needs to be located at least at 50 cm from the scene but should not be as far as 80 cm. The SPAD FOV restricts us to place the sensor at 80 cm from the sphere to capture the entire scene.

6.6 Results

6.6.1 Object masking

The processed range map of the scene is shown in Fig. 6.6 a) where depth ranges from 0 m to 0.6 m. The available size array of the SPAD is 167 x 128 pixels as the bottom rows are not operational. The absolute range of the sphere, monkey head and cube are equal to 20 cm, 30 cm and 40 cm, respectively. This distance corresponds to the distance between the white plane, used for the reference histogram during the TCSPC acquisition, and each imaged object. Fig. 6.6 b) shows an image of the scene obtained with the mobile phone which will be useful to assess the quality of each mask.

By selecting the distance of each object with the range map, a mask can be created on the mobile phone image to select within which PS imaging can be performed. As ex-

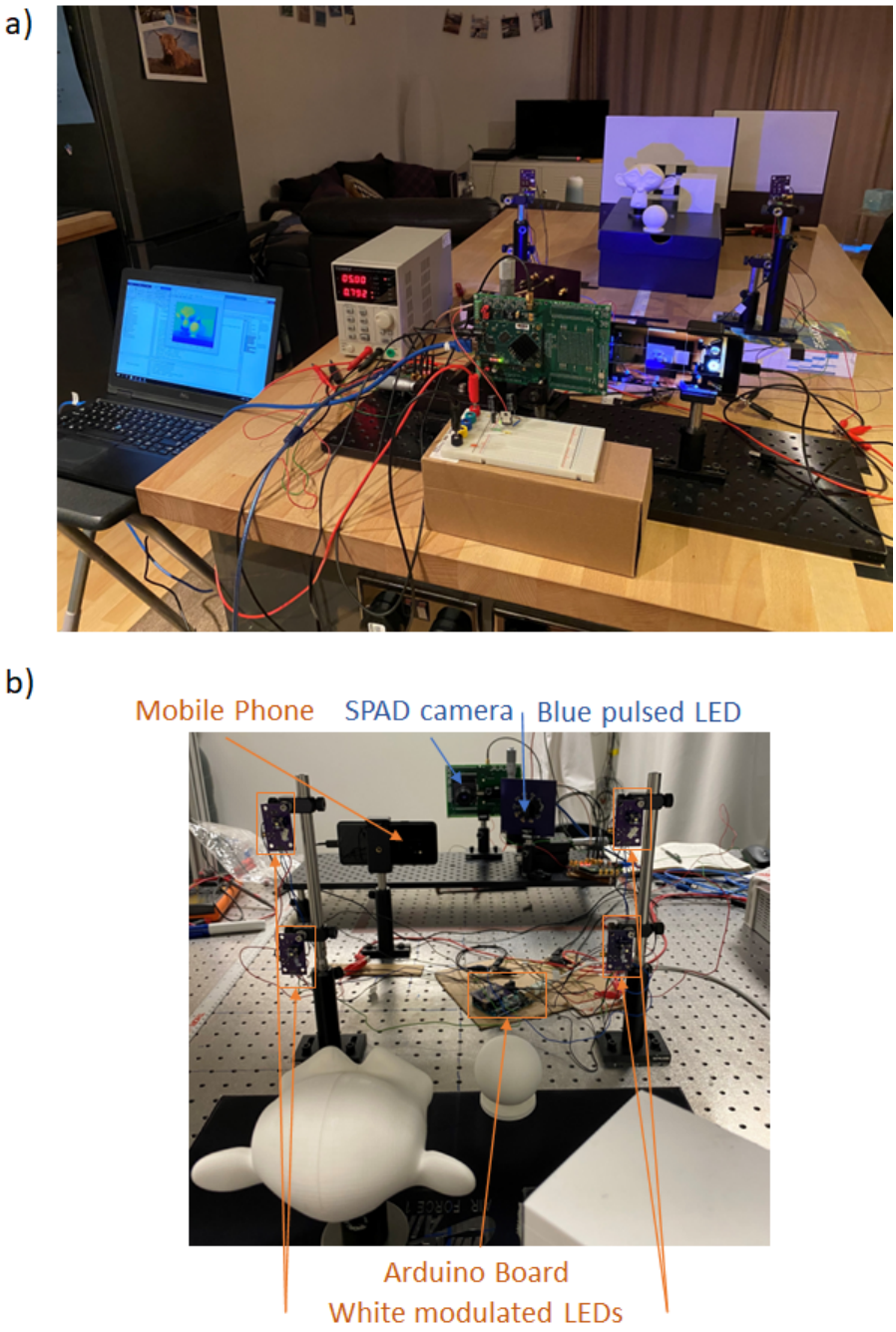


Figure 6.5: a) Working from home: picture of the experimental setup during lockdown period, b) Picture of the experimental set up in the lab.

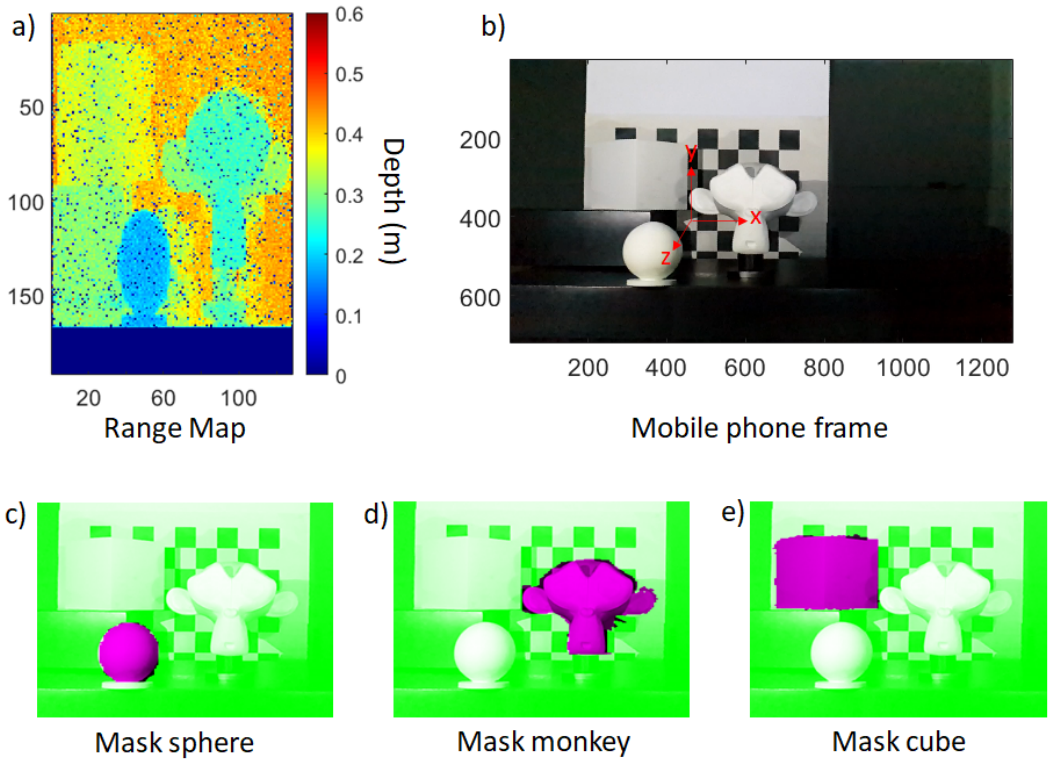


Figure 6.6: a) Range map obtained with the SPAD camera, b) Picture of the scene from the mobile phone, c), d), e) respective masks of the sphere, monkey and cube from the range map superimposed on the mobile phone image.

plained in section 5.2.2, the achieved depth-resolution of the LED-based ToF approach is 3.4 cm. Therefore, to make sure that each object is fully selected, a distance range is determined. For example, to select the monkey, the distance range is set to [0.24, 0.31]. The distance between each object is equal to 10 cm and therefore as long as the selected distance range is below 10 cm then there is no risk of selecting two objects at the same time. In addition, residual noise around each mask has been cleaned using the 'active contour' function from MATLAB^{TM}. The 'active contour' function is a segmentation tool that iterates through a pre-set binary mask to detect the boundary of an object that will be segmented from the background [161]. The corresponding masks of the sphere, the monkey head and the cube are shown in Fig. 6.6 c), d) and e), respectively. Each mask fits each object perfectly on the mobile phone image. Therefore, it is possible to conclude that a cm-scale resolution from ToF imaging is good enough to use it as a masking tool for this situation. Because the resolution is scene-dependant, a cm-scale resolution will be sufficient for different cases such as security scenarios but will not be enough for other cases such as precise range for automative applications.

6.6.2 Surface reconstruction

LED images decoded

The decoded images of the scene obtained with the mobile phone are displayed in Fig. 6.7. The light level shows the different illumination directions of the LEDs placed in an X-shape configuration. Although the LEDs are symmetric to the imaging plan, their brightness is slightly different depending on their position. This can be explained by the possibly imperfect match between the camera integration time and the MEB-FDMA scheme, *i.e.* the integration time might be shorter than the frame duration. Another explanation would be the orientation of the LEDs not being perfectly perpendicular to the XY plane which directs more light to the scene. Nonetheless, the final output images are satisfactory to be used in the determination of the surface normal components in the PS stage.

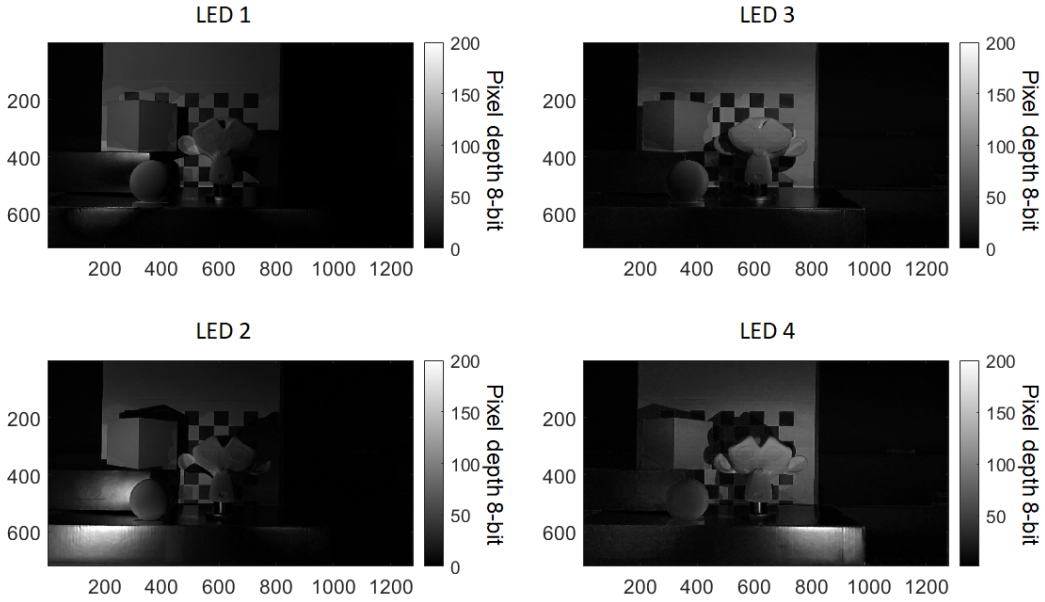


Figure 6.7: Decoded images obtained after demodulation of the recorded frames for LED1, LED2, LED3 and LED4, respectively.

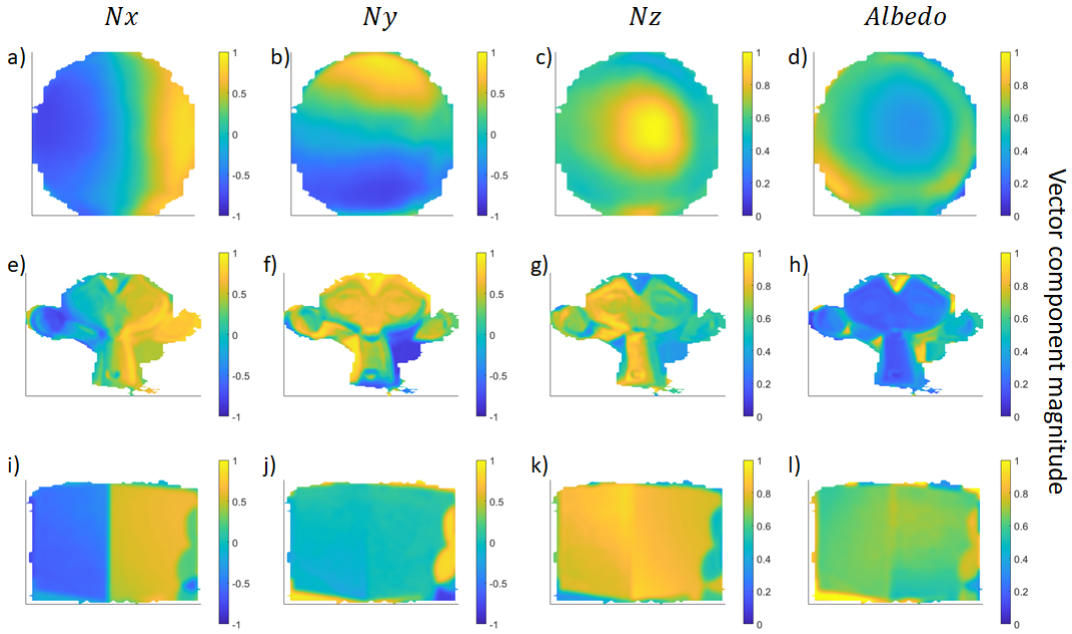


Figure 6.8: Surface normal components and albedo, obtained after running the photometric stereo algorithm, respectively Nx, Ny, Nz and Albedo: a), b), c), d) for the sphere, e), f), g), h) for the monkey head, i), j), k), l) for the cube corner.

Surface normal vectors

Figure 6.8 displays the surface normal components (N_x, N_y, N_z) and the albedo, obtained after the PS processing of the decoded images. The ToF mask has been applied on each object to only select the region of interest. N_x correctly distinguishes left and right facing surfaces, N_y indicates up and down facing surfaces and N_z shows depth variations as the camera is facing the objects. The albedo is normalized and is useful in understanding imperfections in the reconstruction especially on the border of the object.

The cube corner surface normal components, on the right hand side, are impacted by the shadow coming from the monkey head. Fig. 6.7 shows that the shadow mostly comes from the position of LED4. Methods exist to deal with shadows [162, 163]. However, this work focuses on the demonstration of a hybrid technique that combines ToF and PS imaging. To avoid shadows as much as possible, the sphere, the monkey head and the cube are positioned accordingly to the LEDs to minimize the impact of shadows. Although this approach is not ideal, it allows us to keep the same PS process to prove the concept. The impact of shadows on the surface normal clearly shows that this is a parameter that needs to be taken into account in future work. For the moment, it is expected that the shadow area will impact the surface reconstruction.

Surface Reconstruction

Figure 6.9, Fig. 6.10 and Fig. 6.11, respectively, plot the 2.5D reconstruction of the sphere, the monkey head and the cube in a perspective view, a top view, a rendered view and their corresponding ground truth. The rendered view is obtained with Blender$^{\text{TM}}$, a 3D creation open source software introduced previously, where a camera is set at a distance of 10 cm from the imported 2.5D reconstruction. The RMSE error map is plotted for the sphere and the cube.

For the sphere, Fig. 6.9 a), b), c) show a satisfactory reconstruction of the sphere and, as the object is static, only the front can be reconstructed. By comparing the top view with the ground truth, the reconstruction is almost a perfect match. The standard deviation of the surface reconstruction is determined by the RMSE and the

NRMSE [4], see Chapter 4. According to the RMSE error map in Fig. 6.9 d), the error is minimal across the centre of the sphere and ranges up to 6 mm at the bottom which is due to the base that creates a shadow in this area. Nonetheless, most of the error remains under 3 mm. The calculated RMSE over the sphere is equal to 2.06 mm which corresponds to a NRMSE of 4.3 %. This reconstruction error is comparable with the stand-alone PS work in Chapter 4.

For the monkey head, Fig. 6.10 a), b), c) show a 2.5D reconstruction that is slightly tilted on the y-axis. The main reason is due to the LED position with respect to the monkey head. Indeed, in order to keep a consistent analysis across the three objects, the surface normal components are calculated with the LED coordinates defined as in Fig. 6.4. This means that the monkey is considered in the center of the LED coordinates but is actually slightly shifted to the right hand side by a few centimetres. In addition, because of the tilted reconstruction, it is difficult to align the monkey head ground truth with the surface reconstruction. An RMSE calculation on this object would therefore be dominated by the effective tilt induced by the lighting vector offset. In future, compensation of relative object position could be implemented in the recovery algorithm. In this case, the RMSE figure is not represented as it is not a true representation of that figure of merit. Nonetheless, the rendered view of the 2.5D reconstruction shows mm-range details that are reconstructed, including contours of the eyes, the nose and the mouth. By comparing each view with the corresponding ground truth, the face details and shape are reconstructed although some error remains between the ear and the face due to the high z-gradient of the object at this location.

For the cube, Fig. 6.11 a), b), c) show a satisfactory surface reconstruction of the corner with some errors on the bottom right due the shadows which was highlighted before. On the rendered view, an indent is observed and can be explained by the position of the starting point of the Fast Marching reconstruction process, from which the gradient values are integrated in a propagating fashion. At points with a strong gradient variation, the propagation will happen with very different gradient values in different directions and can create an indent on the row or the column of the starting point. However, by comparing the top view with the ground truth, a close match is

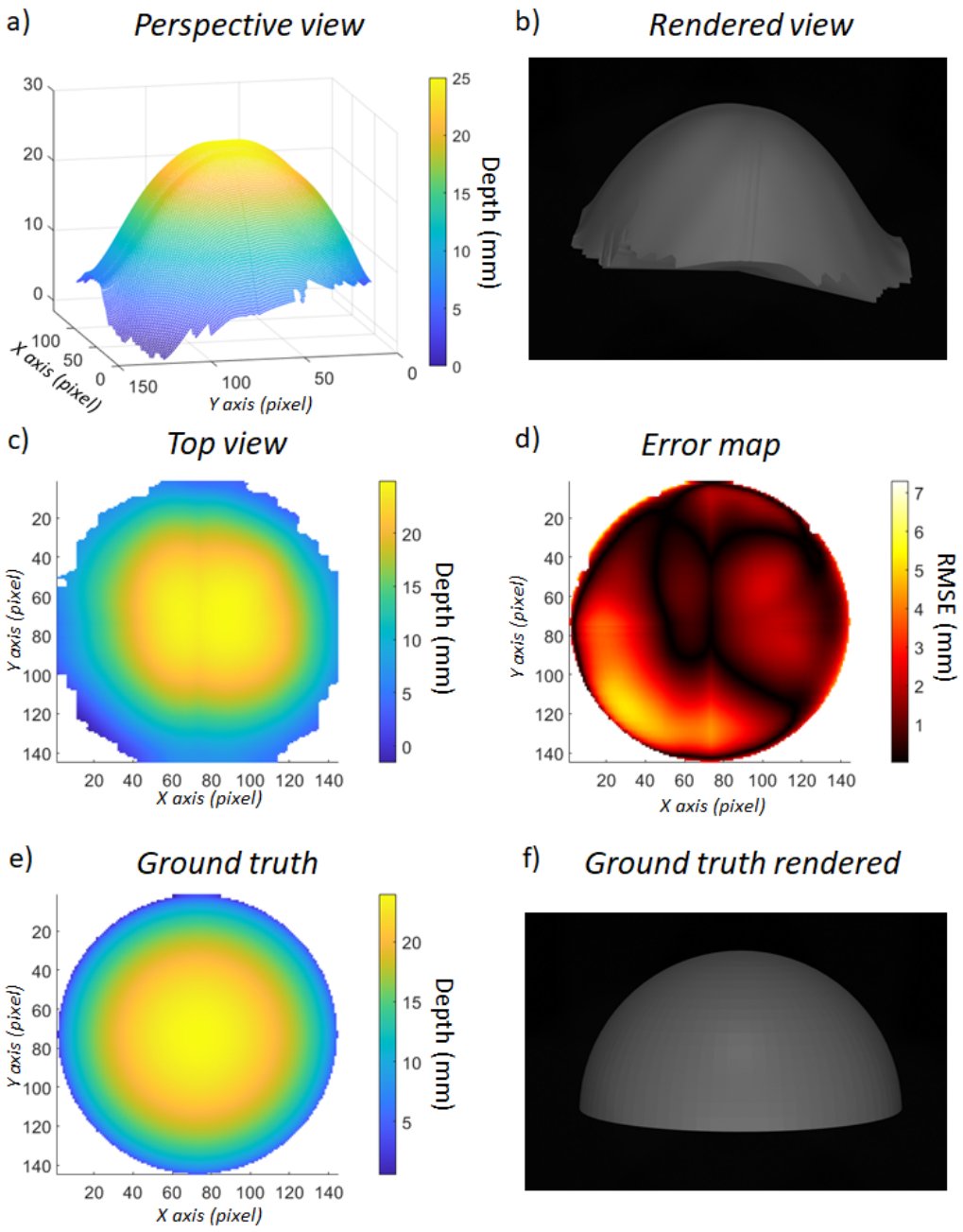


Figure 6.9: 2.5D reconstruction of the sphere. a) Perspective view, b) Rendered view, c) Top view, d) RMSE error map, e) Ground Truth top view and f) Ground Truth rendered view.

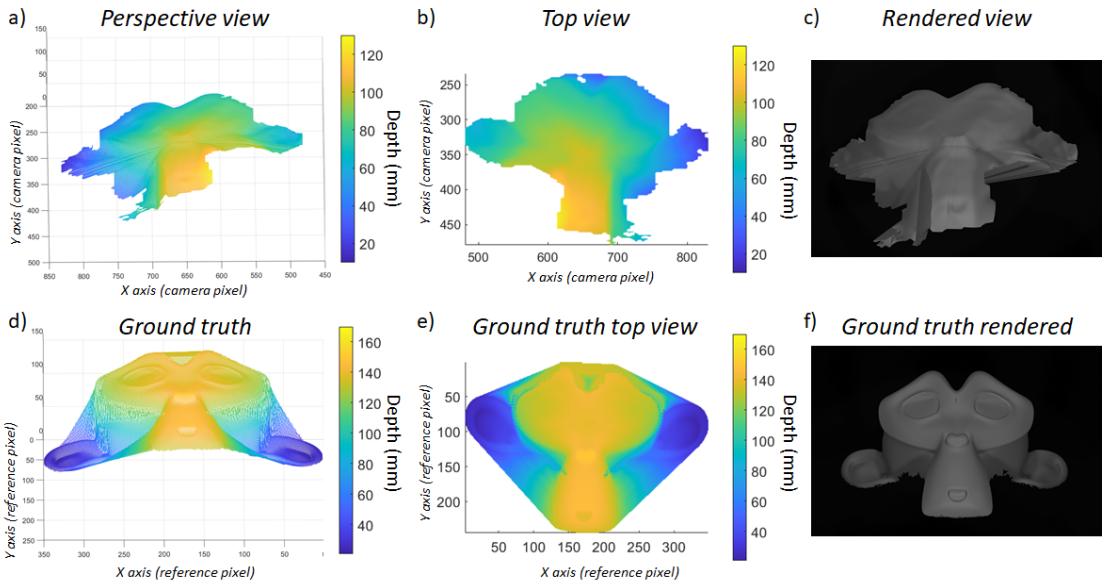


Figure 6.10: 2.5D reconstruction of the monkey head. a) Perspective view, b) Top view, c) Rendered view, d), e), f) Ground truth perspective view, top view and rendered view respectively.

observed. Moreover, when Fig. 6.11 f) is compared with the perspective view, the reconstructed corner shape is similar to the ground truth. The standard deviation of the surface reconstruction is determined by the RMSE and NRMSE, using Eq.4.1 and Eq.4.2 respectively. The RMSE is equal to 3.6 mm with a NRMSE of 4.8 %. The RMSE error map in Fig. 6.11 d) shows that the error is mostly varying between 1 mm and 5 mm and the maximum error is obtained at 16 mm in the shadow area.

The three reconstructions are comparable with the stand-alone PS work [142] which means that the auto-ToF masking does not introduce any additional error and handles the discontinuity issue well. However, there is some difficulty for the algorithm to properly reconstruct the boundary area that is affected by discontinuities and shadows.

6.6.3 Signal to noise ratio

As explained in Chapter 4, the MEB-FDMA modulation scheme used with PS imaging can operate in the presence of additional unmodulated lighting. However, the TCSPC acquisition needs to be performed in a dark room otherwise the detection of the blue pulsed LED by the SPAD sensor would not be accurate. As the calibration needs to be

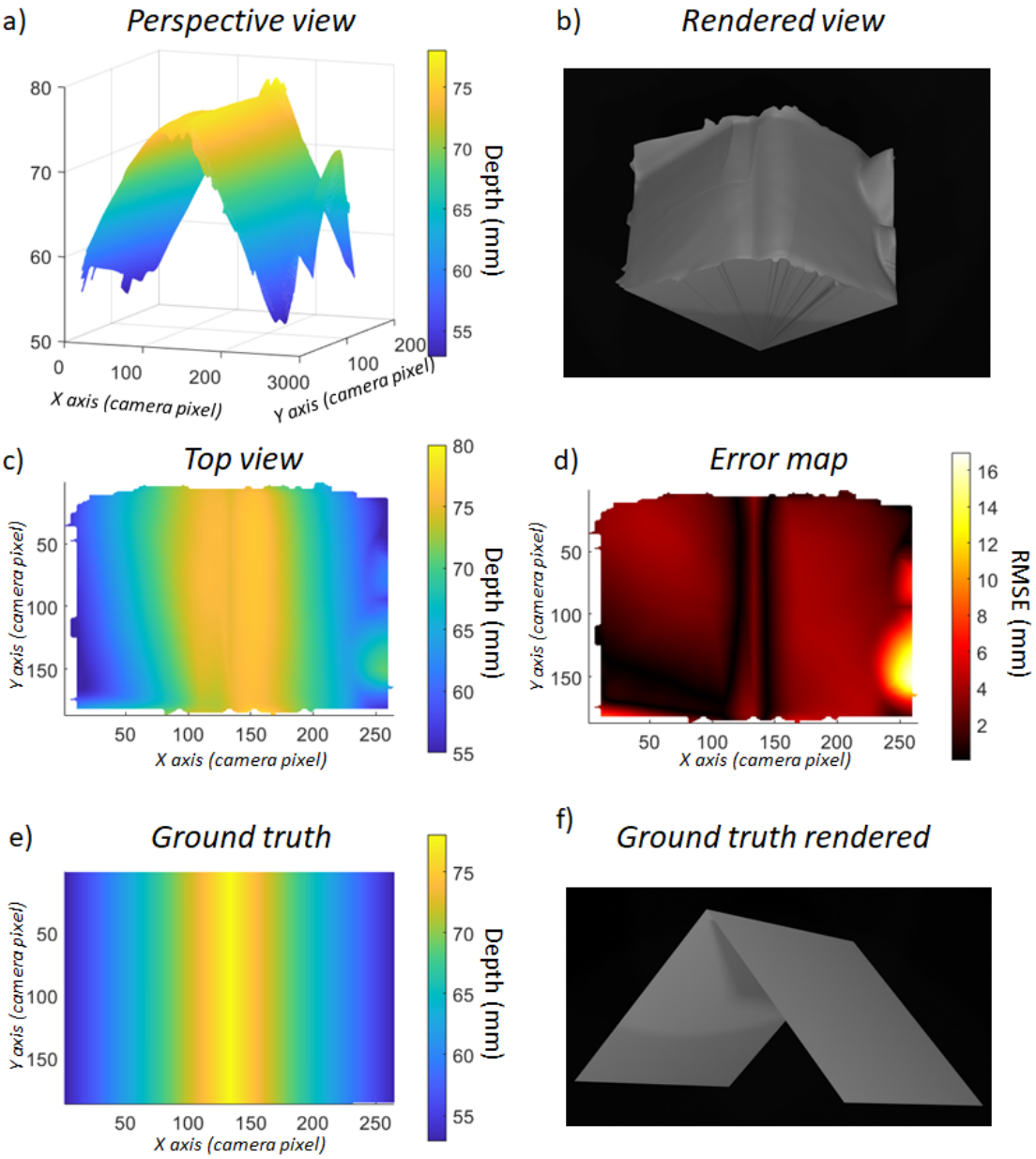


Figure 6.11: 2.5D reconstruction of the cube. a) Perspective view, b) Rendered view, c) Top view, d) RMSE error map, e) Ground Truth top view and f) Ground Truth rendered view.

run only once, it is possible to run the TCSPC one time and then run the PS imaging as many time as necessary as long as both sensors remain static.

To assess the robustness of the PS imaging method with unmodulated background light, the signal to noise (SNR) is measured for the sphere and the cube. The experimental setup remains unchanged, see Fig. 6.4. Similarly to Chapter 4, the ceiling light of the room is illuminated and the brightness of each LED is controlled by an Arduino board. The optical power is measured at the sphere and cube position, which gives different SNR signal for the two objects. The range map is acquired in a dark room before starting the SNR measurement.

Figure 6.12 plots the RMSE error with respect to the signal to noise ratio for the sphere in blue and the cube in red. In both cases, 100% of the surface is reconstructed as the LEDs are not configured in a top-down illumination scheme. Therefore the percentage of the surface reconstructed is not measured.

First of all, the variation of the sphere RMSE curve is between 2 mm and 4 mm with no specific trend across the SNR even in negative value. Futhermore, the sphere error is lower when the SNR equals 4.5 dB and this gives the intuition that high-resolution surface reconstruction can be achieved with LEDs powered with less energy which can fit well with general lighting indoors application.

For the cube, the RMSE curve also does not show a specific trend. However, the RMSE increases up to 8 mm error to then decrease to 4 mm. The increase of the error is not caused by the background light but more likely by the decoded images obtained after decoding the stack of frames acquired with the mobile phone. As the cube is further away by 20 cm from both the LEDs and the mobile phone comparing to the sphere and is impacted by shadows, the surface reconstruction is slightly more challenging, hence the error is higher. Nonetheless, most of the cube reconstruction remains under an RMSE of 6 mm which is in line with results reported in Chapter 4.

6.7 Discussion

Our hybrid strategy of combining ToF with the PS imaging system has proved to be successful. The LED-based ToF method, applied as a masking tool to auto-select

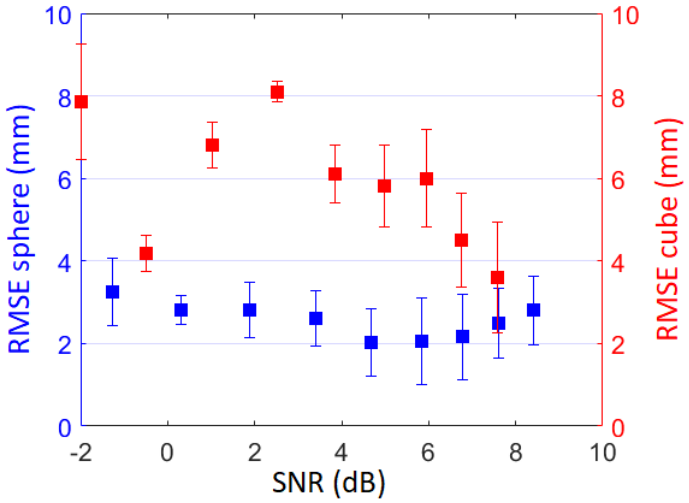


Figure 6.12: Signal to noise results. Graph plotting the RMSE error regarding the signal to noise ratio for the sphere (blue) and the cube (red).

objects within a scene, has been shown to perfectly solve the PS imaging discontinuity issues. The main challenge with the hybrid approach is the calibration of the two different CMOS sensors. By simply using a chequerboard along with a scaling and frame alignment process, it is demonstrated that depth-masks, of 3.4 cm depth-resolution, were obtained with the TCSPC SPAD camera. A high-resolution 2.5D reconstruction of the object as obtained with the PS imaging setup, with an error ranging from 4% to 5%. No more black background was needed with the PS setup. Another advantage of using this predetermined LED-based ToF technique as a masking tool is that no new imaging routine had to be implemented, only the one-time calibration of the sensors was required, plus the PS imaging routine remains unchanged which is therefore time-efficient.

However, to make this hybrid system more time-efficient and more deployable to real-life applications, both cameras need to be bound together. Every time one camera moves while the other one stays still, a calibration is needed. By mounting both cameras together to create one hybrid imaging system, the number of calibrations would be reduced, as the calibration would be required only when the intrinsic parameters of the camera are modified, such as the focus length. With this approach, the en-

tire hybrid system can move within the scene without having to recalibrate it. When mounting the cameras as one solid box, the working distance of the SPAD camera and the commercial camera would be determined in order to define a global FoV. RGB-D camera already exist on the market where a ToF system is added to a commercial RBG camera to improve ToF depth map [74]. Recently, Basler [14] commercialised a ToF camera with the possibility to mount an RGB camera by 3D printing a support that can be attached to the ToF camera. Once the focal length of the RGB camera is set, the one-time calibration runs to determine the corresponding rotation and translation matrices between both cameras. A fusion of a 3D point-cloud and a colour map is then possible which outputs a coloured point-cloud where two objects in the same plane can be distinguished by their colours. A picture of this system is shown in Fig. 6.13 and is a good example of how the hybrid system may look at a future stage. With a rigid mounted and calibrated system, it should be possible to automate mask selection based on depth ranges and continuity of regions in the SPAD image e.g. allowing separate masks for objects in the same plane but laterally separated as well as objects in different planes. With the Basler camera example, the colour image from the smartphone could be used to differentiate objects in the same plane. By simply projecting the 3D information onto the smartphone colour image basis, colour information of the scene could be obtained on top of the depth information acquired with the SPAD sensor. This scenario would enable the hybrid system to be adaptable to any kind of scene by automatically selecting objects by selecting a distance first and then a colour. In addition, for the LED-based ToF system to run in a scene with unmodulated background light and, potentially interferences, the light-pulse emitted wavelength would need to be modified to the near-infrared or other non-visible wavelength.

Nonetheless, the current LED-based ToF system is not suited for depth-map imaging of objects in motion because of its current frame rate and processing time. As explained in the section 5.2.2, the SPAD sensor can potentially reach an acquisition rate of 18.6 kfps with a proper implementation of the FPGA. Physically, there is no fundamental limit on the frame acquisition rate on the transmitter side as the resolution relies on the pulse width of the emitted light pulse. The narrower the pulse

Figure 6.13: Image of a printable solution offered by Basler to obtain a colour 3D point cloud by mounting a RGB camera with a ToF camera [14].

width, the higher the depth-resolution. As long as each pulse can be differentiated in the cross-correlation process, then only the electronic driver capacity can limit the transmission rate. The sensor remains the main limitation in the ToF frame acquisition as the dead time and the fill-factor are the two parameters limiting the pulse repetition and the sensor efficiency, respectively. Therefore, a trade-off has to be found between the achievable sensor efficiency (including acquisition frame rate and fill-factor), the achievable light pulse width and the signal processing approach (including TCSPC and cross-correlation process) in order to achieve a fast acquisition of depth-map scenes at high-resolution. For example, in theory, if the SPAD acquisition frame rate is set at 18.6 kfps to acquire 10,000 TCSPC frames, hence the same number of frames acquired to build up the histogram, then the acquisition time would be equal to 0.5 s which is 40 times faster. In this instance, the acquisition frame rate would meet the requirement to possibly achieve real-time imaging when considering an adequate signal processing approach. Indeed, the histogram build-up and the calculation of the range-map are parameters to consider with real-time imaging. Because of the low-photon count nature of the ToF system, Poissonian noise can become dominant if not enough frames are acquired [62]. As the frame acquisition increases, the number of detected photons decreases which can lead to a higher number of frames acquired, hence a trade-off needs to be found to obtain the most efficient approach. Currently, the depth-map is built in 5 minutes (acquisition and data processing) using Matlab on a laptop and is mostly dependent on the cross-correlation process between the scene and the reference.

To alleviate the computational cost, future work will focus on the fast computational approach of LED-based ToF real-time imaging.

Another issue raised in this chapter is the impact of shadows on surface normal calculation and the surface reconstruction. For both the cube and the monkey head, in the mouth area, the corresponding 2.5D reconstructions were impacted by some shadows and resulted in lower-resolution reconstruction. Some methods currently exist where shadow areas are detected and where the PS method is adjusted to the number of illumination directions made available in the affected area [162, 163]. These methods should be readily applicable to the hybrid approach, requiring a modest modification of the image processing routine.

6.8 Conclusion

In this chapter, the results demonstrate that ToF can be used as a masking tool to overcome discontinuous issues of the PS method, using illumination with LEDs and CMOS-based image detectors in both cameras. The experimental work focused on demonstrating a simple combination technique that works well without having to change the PS imaging process. Despite the cm-scale depth-resolution of the LED-based ToF method, accurate masks were produced to then achieve high-resolution surface reconstruction with the PS imaging. The RMS error ranges between 4% and 5% for an object auto-selected from a scene imaged at a distance of 50 cm to 70 cm. Future work will focus on adapting the imaging process to take shadows into considerations.

Chapter 7

A Photometric Stereo dataset for deep-learning application

Chapter 6 dealt with the discontinuity issues by using the LED-based ToF and PS hybrid imaging system and achieved high-resolution of auto-selected objects. Now, this chapter will investigate the possibility to implement deep-learning in order to improve the computational time of the PS reconstruction procedure developed in Chapter 4 with a view to achieve real-time imaging as mentioned in Chapter 5. The powerful tool of deep learning such as a Convolutional Neural Network (CNN) could be trained to estimate the depth map of different kind of objects as compared with computationally intensive spatial integration method, *i.e.* the Fast Marching method. The main contribution here is the generation of a dataset that would allow the implementation of such deep-learning capacity in the future. A dataset that contains different illumination directions onto a range of low spatial frequencies objects with ground truth shape as an output has not been done to date. Most of the dataset available in PS imaging are focused on the generation of surface normal but not the final topography of the object. The method is therefore demonstrated here and the dataset made available in the thesis repository [19]. In addition, this chapter explains why a deep learning approach is a good candidate to solve the outstanding issues in image recovery from photometric stereo datasets such as fast computation PS, light calibration, specular reflection and shadows.

7.1 Motivation

As discussed in Chapter 5, the end goal of the 3D imaging system is to achieve real-time reconstruction of scenes in order to successfully reconstruct real-time scenarios. Based on all the previous chapters, it is clear that real-time 3D imaging is a challenging task especially when other issues arise such as light calibration and shadows. In addition, the assumption made in Chapter 4 on Lambertian reflection [39] will not be applicable in most cases as common objects incorporate both diffuse and specular reflective elements. To make the depth reconstruction system a tractable problem the application space is restricted in the first instance to diffuse reflective objects. The consideration of uncalibrated PS imaging, shadows and specular reflections is not the first priority in this work. However, these features could easily be added to the real-time reconstruction problem if a deep learning approach is selected.

Recently, with the great success of deep learning in various computer vision tasks [16], deep learning based methods have been introduced to deal with the different issues encountered with PS imaging. Non-Lambertian surface reflection [15, 49, 51, 164] and uncalibrated PS imaging [15, 51, 165, 166] are the main topics covered in the literature at the moment. Moreover, some works also report successes in using a trained neural network to achieve 3D PS-based face recognition [147, 166, 167], for example. Other important features are covered with the deep learning approach such as the near-field PS problem [168], the non-convex shape [169] and the single snapshot problem [166], where the latter can also be dealt by multi-spectral PS imaging [167, 170]. The deep learning research area in PS imaging has clearly been well investigated and shows interest in combining PS features to create more robust models for reconstructing the shape of real world objects which could then be applied to different applications including face recognition. Moreover, an interesting, and useful, observation is made in the work of Guanying Chen *et al.* [15, 51, 165] as there both the non-Lambertian reflectance map and the calibration issues are successfully tackled. These results are highly encouraging for the work demonstrated here as the goal is to have a PS imaging setup that can be mainly deployed in building infrastructures for real-time video surveillance applications

which will contain various materials, objects shapes and lighting conditions.

In this chapter, a brief overview on deep learning will introduce the important ideas in deep-learning such as under fitting and over fitting issues, along with the training of a network and its validation. In addition, an explanation of the signification of a CNN will be given along with the reasons why it allows for a speed up in the data processing. Then a general overview on the dataset found in the literature will introduce the generation of a bespoke dataset for PS imaging. A main task in creating a CNN is to have a large sample dataset to train the model with. The major part of the work reported here is in the creation and assessment of a suitable training dataset. For the model to retrofit the experimental setup used in Chapter 4, the computational rendering tool Blender has been chosen as it re-creates similar working conditions to the top-down illumination scenario with the camera facing the objects. Blender has multiple settings available in term of rendering (camera, pixel resolutions, etc.), object textures and light sources. For the rendered dataset to be compatible with the real captured image data, the rendering features can be tuned to reflect the real setup.

At last, an introductory work on the implementation of a deep-learning network will be presented based on the CNN developed by G. Chen [15], who made it accessible on Github, and the goal is to adapt the input/output dataset to estimate the depth map instead of the normal map. Fig. 7.1 displays the Photometric Stereo Fully Convolutional Neural Network (PS-FCN) [15] which is composed of three different layers, namely the *shared-weight feature extractor*, *the fusion layer* and *the normal regression network*. The fundamentals of the network architecture remain the same although, by optimising the CNN on the ground truth depth map, the CNN should learn to estimate the depth map and not the surface normal map. Besides achieving a fast PS computation, it would be possible to train a self-calibrating CNN by removing the light directions from the input of the PS-FCN as shown in Fig. 7.1, which in theory makes this method compatible with the important use case of uncalibrated PS imaging. The self-calibrating CNN has not been tested in this thesis but is explained for possible future application of the CNN.

This chapter will conclude on the factors that need to be considered when estimating

the computational time response of a trained and optimised model, such as the image resolution, the computational complexity of the model and the processor it is running on [16]. In a particular case with an VGA image resolution and a CNN model of two CNN networks, for example, the processing time would be less than a second. As the implementation of the CNN is only introductory and is made to draw a path for future work, a conclusion on whether the CNN model increases the computational time has not yet been drawn and therefore the steps needed to finish the implementation of the CNN will be explained.

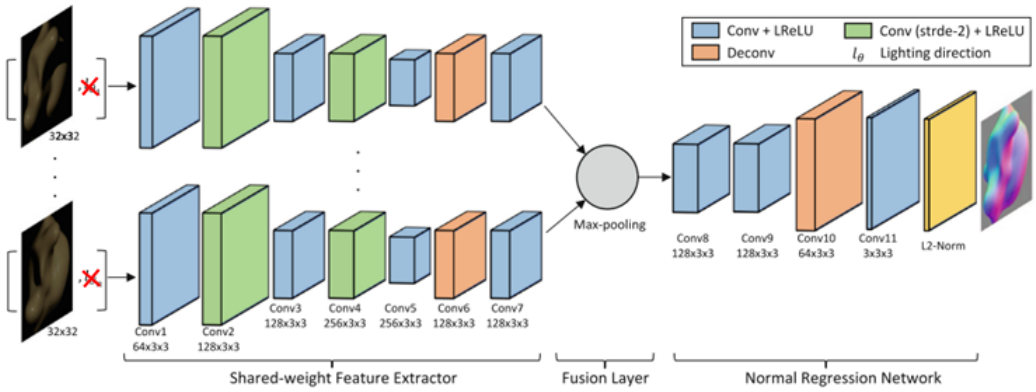


Figure 7.1: Network architecture of PS-FCN [15].

7.2 Deep learning

A quote to explain what machine learning is would be the following: "An Artificial Intelligence system needs the ability to acquire its own knowledge by extracting patterns from raw data." [16]. In simple words, machine learning enables computers to tackle problems involving knowledge of the world and to make decisions. Deep learning is a deeper machine learning algorithm, *i.e.* an algorithm that contains deeper layers to execute complicated tasks, such as face recognition, see Fig. 7.2 a). A deep learning algorithm enables the computer to build complex concepts out of simpler concepts [16]. For example, to detect a car, the algorithm will first look for the corners and contours, then the edges and then the shape of the car. Deep learning relies on two perspectives.

First, learning the right representation for the data as a good representation will have an impact on the performance of the algorithm. Second, depth enables the computer to learn a multistep computer program. Basically, each layer of the representation can be thought as the state of the computer's memory after executing another set of instructions in parallel. Technically, networks with greater depth can execute more instructions in sequence [16]. Scale drives the deep learning process, the more data the more efficient a large neural network will be compared to a small neural network for the same amount of data, see Fig. 7.2 b).

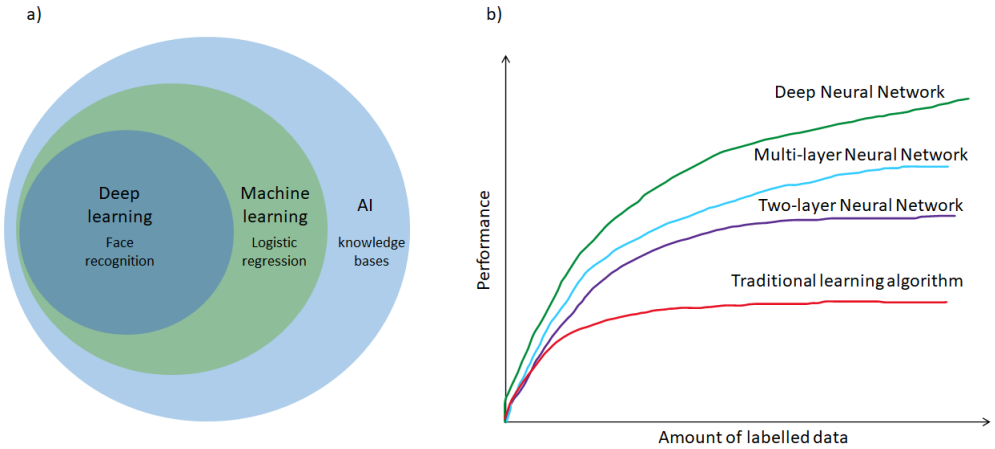


Figure 7.2: a) Venn diagram showing how deep learning is a kind of machine learning, which is used for many approaches to AI [16]. b) Hand drawn schematic showing the difference in performance of the different neural networks regarding the amount of labelled data used for the training.

7.2.1 How to train a deep neural network

A neural network can be supervised or unsupervised, which in the latter case means that no output data is used to classify the input features. Unsupervised learning is out of scope in this work and therefore the introductory work will focus on how a supervised learning neural network is trained and validated. In addition, the solutions on fixing different issues a model training can experience will be developed. Fig. 7.3 a) shows a schematic representation of a neural network with the input, the hidden layers and the output.

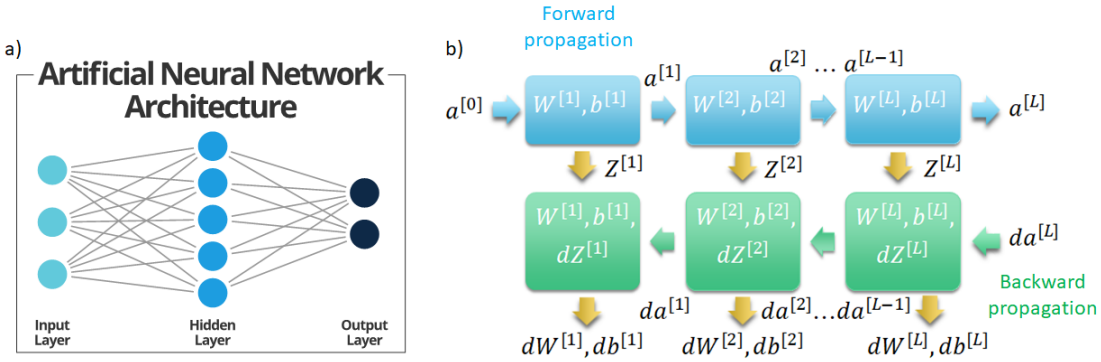


Figure 7.3: a) Schematic of an artificial neural network. b) Building blocks of a deep neural network to illustrate the forward and backward propagation.

To train a neural network, a few steps are required. First, the structure of the model needs to be defined, for example the number of input features and the size of the dataset. Then, the model's parameters have to be initialised; in Fig. 7.3 the parameters are the weight of each hidden layer W and b. The third step consists of the training of the model through a defined number of epochs, *i.e.* number of iterations that the algorithm will iterate forward and backward through the neural network model. As shown in Fig. 7.3 b), the first step is the calculation through the blue blocks, called forward propagation, of the prediction value a which will then be used in the loss function to determine the error between the prediction and the true output value. In order to minimise the loss function, the gradient of each hidden layer is determined through back propagation (green block) to then update the model's parameters W and b. This step is called gradient descent which is an optimization algorithm used to minimize the loss function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient. After a few epochs, the gradient descent will reach a minimum, as shown in Fig. 7.4 a). Fig. 7.4 b) shows that the choice of the learning rate is important for the gradient descent to work. The learning rate determines how rapidly the parameters are updated. If the learning rate is too high then the optimal value might "overshoot" and if the learning rate is too small then the convergence will be reached after a lot of iterations, which will make the training inefficient. This quick overview gives a rough idea of how a model is trained, and more

information and details can be found in e.g. [16].

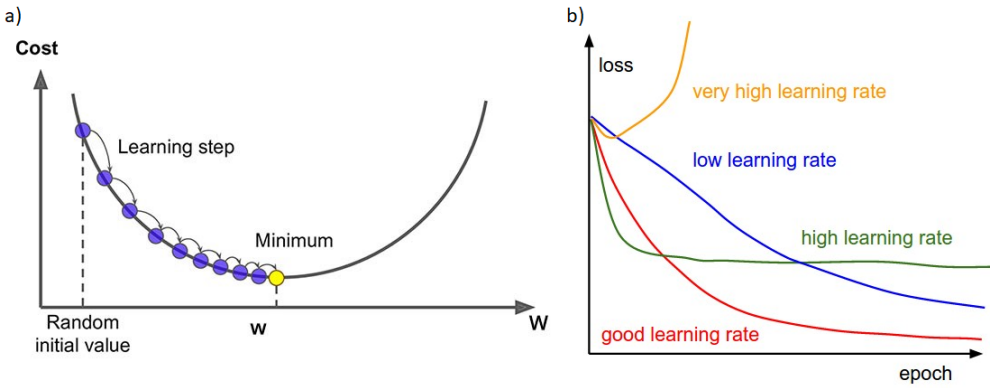


Figure 7.4: a) Gradient descent in 1D which represents the minimisation of the cost function, also called the loss function, by updating the parameters with a learning step that gradually decreases. b) Curve representing the evolution of the loss function for different learning rate value regarding the number of epochs when training a neural network model.

The learning rate is one hyperparameter that can be tuned during the training. Nonetheless, despite a good learning rate, other features need to be considered to avoid any bias in the training. As shown in Fig. 7.5, the prediction error can result in different performance regarding the dataset: underfit, optimal or overfit. An undefitting model highlights high bias issues which means that the model is not complex enough to learn features from the dataset and it results in a high prediction error in both the training set and the test set. An overfitting issue is a high variance problem and results in a model that is too complex to do well on other datasets than the training set, hence gives a high prediction error on the test set only. The model is considered optimal when both the training and the test dataset have the same order of magnitude. In practice, this means that the model has learned well on the training dataset and can be adapted to another distribution set of data. To fix an underfitted model, a solution is to even create a deeper network to increase its complexity or to train the model for longer. Once the bias is reduced, and the model experienced a variance issue then the solution is to have more data to train it on or use a regularisation approach. Stated briefly, the regularisation includes different methods that will decrease the variance, for example, by smoothing the decision boundary with an L_2 regularisation, a dropout

approach, which will remove some connection between hidden layer nodes, or an early stop [16]. A model is validated once the fit is optimal and the model easily deployable to new sets of data.

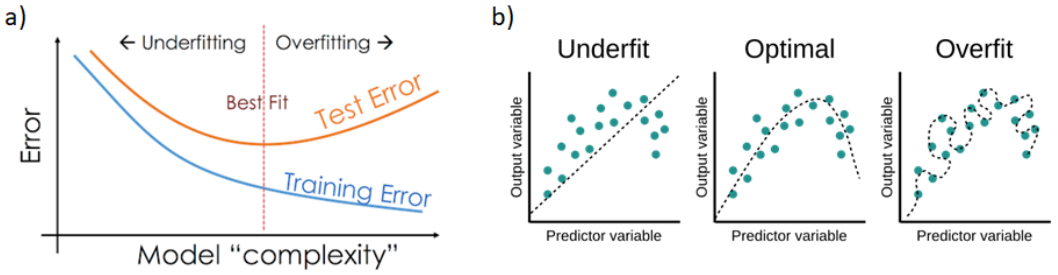


Figure 7.5: a) Curve representing the prediction error regarding the model complexity for the training dataset and the test dataset. b) (Left) A linear function fit to the data suffers from underfitting and cannot capture the curvature that is present in the data. (Centre) A quadratic function fit to the data generalises well to unseen points. (Right) A higher polynomial degree curve suffers from overfitting.

A good model relies on the quality and the distribution of the dataset. A well-chosen dataset can considerably improve the learning as it can possibly deal with overfitting issues [16]. Distribution, here, means the source of the data. For example with an image-base dataset: is the image gathered from internet, from a computer vision tool or from a phone? Has the model been trained on data gathered during summer only? In the case of an animal recognition application, will the model detect the animal during winter time? In some application, it is highly recommended to have a training set distribution that differs from the test one to reduce overfitting problems. The only rule is that the development and the test dataset need to have the same distribution [16]. As an example, when considering a cat detection application for mobile devices, an easy/fast way to obtain a training set would be to gather cat images from the web. However, when the trained model is used on the phone would it generalise well to this captured photograph? By having a development set that contains cat images directly taken from a phone, the model would first learn to recognise a cat from web pictures but then parameters will be tuned to also recognise a cat on phone pictures. In this situation the model will not overfit on the web based images. Some more information on mismatched training can be found here [171, 172].

7.2.2 Convolutional Neural Network

A CNN is a specialised kind of neural network for processing data that has a known grid-like topology, such as an image which is represented as a 2D grid of pixels [16]. CNN uses convolution instead of general matrix multiplication to speed up the computation time and to work with inputs of variable size. Convolution leverages three important ideas that can help improve a machine learning system by incorporating sparse interactions, parameter sharing and equivariant representation. In a traditional neural network, the interaction between each input unit and each output unit is described by a matrix multiplication made of unknown parameters. As shown in Fig. 7.6 every output unit interacts with every input unit. While with the convolution, a kernel filter of size k is used to create sparse interactions. This is important in image processing as the input image might have thousands of pixels, but to detect meaningful features such as edges, a kernel of only tens or hundreds of pixels is needed. Therefore, fewer parameters are stored which both reduces the memory requirements of the model and improves its statistical efficiency [16]. In addition to the sparse interaction, parameter sharing, used by the convolution operation, dramatically decreases the statistical efficiency compared to a dense matrix multiplication as it only learns one set of parameters per location, which reduces the storage requirement of the model to k parameters [16].

Another important feature of a CNN that speeds up the data processing of images is a pooling layer. A pooling function replaces the output of the net at a certain location with a summary statistic of the nearby outputs, see the example with the average and the maximum pooling in Fig. 7.8. It also helps to make the representation approximately invariant to small translation of the input. Because pooling summarises the response over a whole neighbourhood, it is possible to use fewer pooling units than detector units by applying statistics for pooling regions spaced by k (the size of the kernel). Hence, the next layer has k times fewer inputs to process which improves the computational efficiency [16]. Importantly, pooling is an essential tool for handling inputs of different size. For example, to classify images of variable size, the input to the classification layer must have a fixed size. By varying the the size of an offset between pooling regions, the classification layer will always receive the same number of

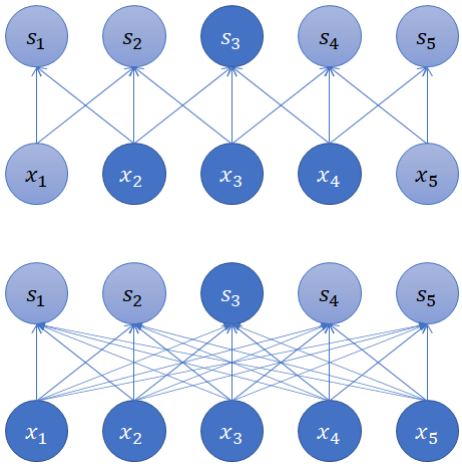


Figure 7.6: Sparse connectivity. One output unit s_3, and the input units in x that affect this unit are highlighted. These units are known as the receptive field of s_3. (Top) When s is formed by convolution with q kernel of width 3, only three inputs affect s_3. (Bottom) When s is formed by matrix multiplication, connectivity is no longer sparse, so all the inputs affect s_3 [16].

summary statistics regardless of the input size.

7.3 Dataset acquisition

When using a CNN, the task of generating a dataset rapidly and efficiently is a challenge because of the amount of data required, which can be on the order of a million samples. A few practical parameters need to be considered in the first place, namely the storage, the diversity of the dataset, the resolution of the images, the software or the source that will generate the dataset, and the type of processor available to handle the computation - for example CPU or GPU hardware. Another aspect to consider is whether to choose to have the same data distribution across the training set and the test set. As explained in the deep learning section above is it better to have a bespoke training dataset or one that will generalise well to different data? What is the target/goal of the CNN model which will then decide on the dataset format?

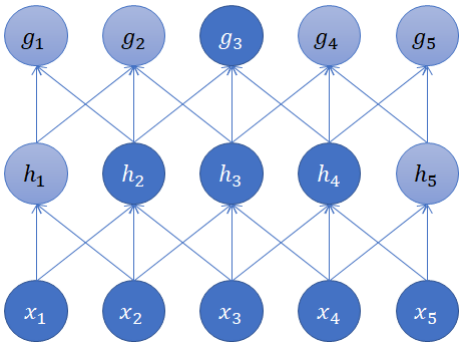


Figure 7.7: The receptive field of the units in the deeper layers of a convolutional network is larger than the receptive field of the units in the shallow layers. This effect increases if the network includes architectural features like strided convolution or pooling. This means than even though direct connections in a convolutional net are very sparse, units in the deeper layers can be indirectly connected to all or most of the input images [16].

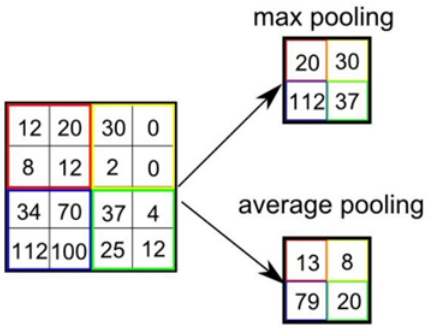


Figure 7.8: Maximum and average pooling.

7.3.1 Dataset in the literature

In this work, the goal is to retrieve the 2.5D shape of an object from the four images obtained after decoding the video with the MEB-FDMA decoding matrix from Chapter 4. As discussed earlier, the experimental top-down illumination setup is reproduced and is similar to Chapter 4 to train the CNN. How can thousands of images be generated for this task when the images are not already available?

Some datasets are made available by research groups, such as the "blobby shape" dataset [173], the "DiliGenT" [174] or the "SilNet" [18]. The blobby dataset consists of 100 images: ten blobby shapes rendered in ten natural lighting environments. The dataset also includes OBJ files, normal maps, masks, and scripts for rendering the images. About 16,000 OBJ files are made available and they represent 3D objects with randomly generated surface topologies of low-spatial frequency. The DiliGenT dataset is a benchmark set designed for PS imaging which contains calibrated directional lighting, objects of general reflectance, and ground truth surface normal for orthographic projection and single-view setup. DiliGenT also provides a photometric stereo taxonomy emphasizing non-Lambertian and uncalibrated methods. Finally, the SilNet both provides the rendered images and their corresponding silhouettes in an obj file, which are considered as synthetic datasets. In [18], Wiles and Zisserman explains that SilNet is also a deep-learning architecture that can generate silhouettes in new viewpoints.

Although these synthetic datasets are not usable as such and are not in a ready to use state for the specific task, the blobby shape dataset is interesting as it contains 3D object files that can easily be imported in Blender [146]. As mentioned in the previous chapters, Blender is a free, open source, ray tracing based rendering software which can support both CPU and GPU. Another interesting feature of Blender is its Python based script rendering operation. In addition, Blender has a specific library called "bpy" which makes it easy to import objects within the scene and to modify it (scale, render level, etc., ...) by applying modifiers as one can manually do on the layout view. In general, computing thousands of images might take a few days on a desktop computer. Based on this, the generation of thousands of rendered images from imported 3D objects will clearly take longer. By scripting the entire process in Python,

the bespoke data generation will be time consuming but it will run without having to intervene. To simplify the training of the CNN, it is decided to only use low spatial frequency shape with the rendering of images from the blobby shape at the beginning. In the future, it would a good idea to add more complexity to the model by using the sculpture dataset, from SilNet [18], which will contain higher spatial frequency.

7.3.2 Blender rendering

To run the synthetic dataset acquisition in the most efficient way, a pre-build setup was build and set to not be modified in the Python script, except for switching on and off the light sources. As a start, the dataset represents the experimental setup that is described in Chapter 4, therefore four illumination sources are created in the top-down scenario and a camera is placed in front of the object. Fig. 7.9 shows a picture of the Blender setup in three different views, perspective, top view and side view. The black background represents the xy plane and is useful in avoiding the back reflection of the light from the background to the camera. The z axis is directed toward the camera. The LEDs are modelled as white disk shape area sources. A fifth LED is added above the object in the same direction as the camera to easily create a mask of the rendered object, which can be useful during the CNN training. As this is a model, the shape of LED 5 is not seen by the camera and only illuminates the object when activated. The entire setup is scaled in a way that the rendered images are similar to the ones obtained in Chapter 4. Table 7.1 details the parameters selected for the camera option, the render engine and the resolution of the rendered image.

Table 7.1: **Summary of the main parameters used for the Blender model.**

Camera type	**Render engine**	**Resolution**
Orthographic	Cycle	X: 480 pixels
	support GPU compute	Y: 270 pixels
	Integrator path ray tracing	PNG 8bit depth
	Render 100	Compression 15 %

The Python script imports each blobby shape object and applies some modifiers to make the object smooth and less heavy memory-wise by reducing the number of

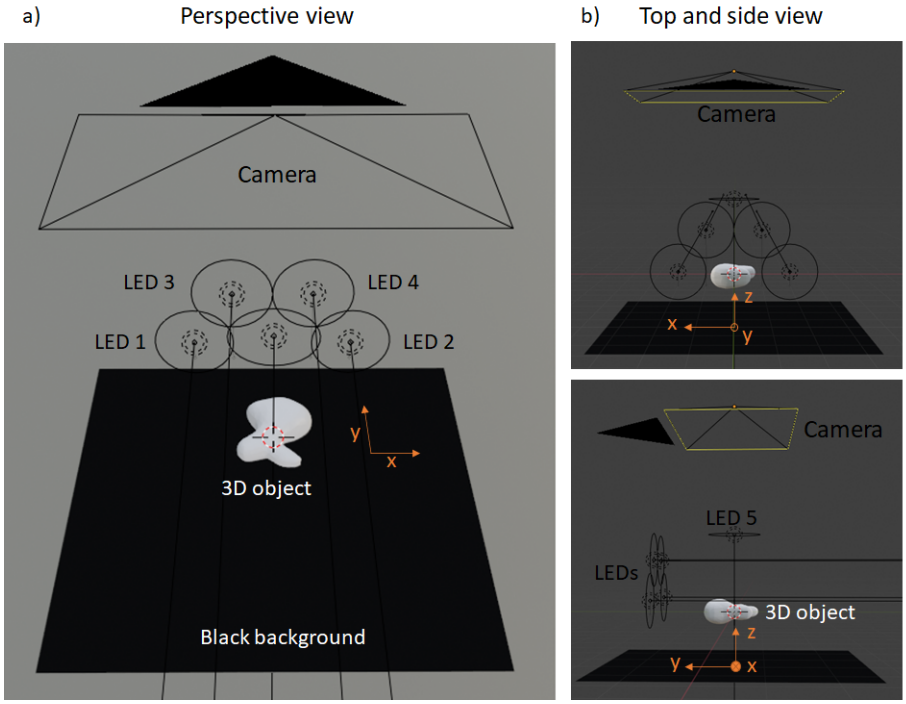


Figure 7.9: Blender model for data set acquisition, a) perspective view, b) top view and side view. The camera is represented by the rectangle, the LEDs by the disks and the blobby shape object is in the middle. A black background is here to remove the back reflections.

vertices that the OBJ file contained. Then, the script switches on one light source at a time to render an image that is directly saved in the dataset folder. After rendering the four images, the LED 5 is switched on and a threshold is applied to the rendered image to create a binary mask of the object. This is also saved in the same folder and will be used later on in the CNN. Moreover, in case the light direction of each LED is needed to improve the learning, the lighting vectors are calculated for the X, Y, Z axis and saved in a text file. Before removing the imported object, an STL file is created to save the modifications applied to the object, only the object being selected and saved in the STL. This is file is then processed in Python to be converted into an array which represents the object depth. By creating an output of the same image input resolution, which represents the true depth, the output data images that will be used in the CNN for supervised learning are created. After processing one object, the Blender script removes it from the scene and imports the next one.

The blobby shape dataset contains 15,726 different objects for a memory size of 4.16 GB. Four images of resolution 480x270 pixels are generated per object along with a binary mask and an STL output file. The STL file memory is relatively important, about 12 MB per object, which requires at least 200 GB of memory for the STL file dataset to be saved on the disk. As explained in the paragraph above, the STL file is processed in a Python program to generate the ground truth depth array. However, this task is done after generating all the STL files. Hence, to reduce the memory load of the whole dataset (inputs and outputs), the ground truth arrays are not saved as an image file but are stacked in an H5PY file which considerably reduces the memory size of the dataset. A text file keeps track of the data saved in each H5PY file so that each ground truth output can correctly be paired with the input images. Table. 7.2 shows an overview of the organisation of the dataset with the file type and the memory requirement. The data acquisition ran discontinuously on a desktop computer for a total of two weeks. It processed the full 15,726 3D objects to obtain the rendered images, as detailed in table 7.2. As a result the final dataset contains four input images, a binary mask and a ground truth depth array per 3D blobby shape, which in total represents a dataset of 8.8 GB.

Table 7.2: **Overview of the generated blobby shape dataset.**

	PS images	**Binary mask**	**Lighting vectors**	**Ground truth**
Global dataset	62,904 images	15,726 files	15,726 files	16 files
Per 3D object	4 images	1 file	1 file	1 numpy array stacked in one HDF5 file of 1,000 outputs
Global memory size	4.03 Go	62,9 Mo	2.3 Mo	3.57 Go
File format	png	png	txt	HDF5

7.3.3 Rendered synthetic dataset

Fig. 7.10 displays a selection of the input images that will be used in the CNN for training and gives an outlook of the different kind of shape that are contained in the blobby shape dataset. By choosing a randomly generated shape dataset that covers a specific range of spatial frequencies, the training set can cover a reasonable real world scenario. If the training dataset were to contain repetitive features then the training would become biased and result in over fitting issues. At the moment, it is preferred to simplify the training by only considering low-spatial frequency features, but, in the future, it will be necessary to add higher spatial frequencies in the training set by adding the sculpture dataset from SilNet [18] for example. The shading of the four illumination directions can be observed on the objects, which is made more noticeable in Fig. 7.12. Fig. 7.10 also shows the way each image has been named to then be easily retrieved. In a situation where a larger dataset would be required, it would be straightforward to implement a script on Blender that would rotate the object to generate a higher number of rendered images. By keeping track of the ground truth depth with the STL file, it would be easy to generate a double amount of datasets. Then a trade-off between the storage memory available, the computer GPU that can support the rendering and the accuracy of the trained CNN would need to be considered. This is not something that will be discussed here although it is clear that storage memory can rapidly become an issue.

Once the dataset is generated, a decision has to be made on the way to separate it into a training set and a development set, which is called here a test set. Four images per singular shape were rendered, the blobby shape dataset containing 15,726 singular shapes which represents a total of 62,904 images. Following the guidelines of deep

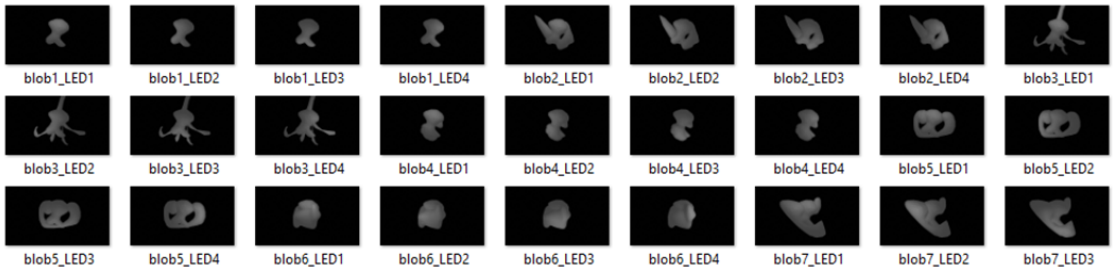


Figure 7.10: Selection of rendered images using the Blender script. Four images are rendered for each blobby shape, one image for each illumination direction.

learning courses [16], the dataset is below a million data therefore the split between training and test set will be 80:20, respectively. Fig. 7.11 shows an example of the superposition of the input image with the ground truth depth map to make sure that they can be superposed perfectly which is important to obtain an accurate learning.

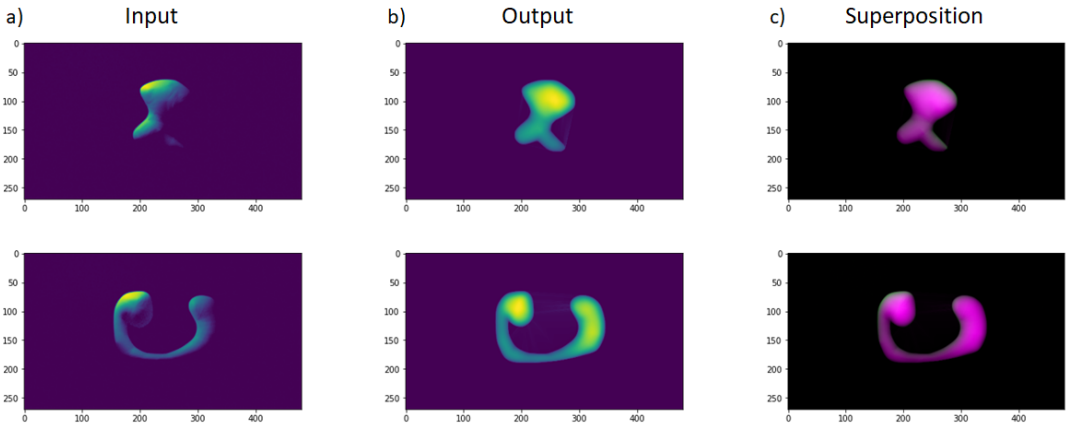


Figure 7.11: a) Example of an input image for two different objects, b) the corresponding output depth map and c) the superposition of the two to confirm a good match for accurate learning.

7.4 Depth estimation convolutional neural network

To get started on the architecture of the CNN, the PS-FCN made available by G.Chen [15, 51] has been used and adapted as a depth estimation CNN instead of a surface normal estimation network. The idea is that by simply modifying the output of the supervised learning network into a ground truth depth map, the CNN should learn to estimate

the depth of the input images instead of its surface normal. The adapted PS-FCN is a multi-input-single-output network composed of three components, namely a *shared-weight feature extractor* for extracting feature representation from the input images, a *fusion layer* for aggregating features from multiple images, and a *normal regression network* [15, 51], which is modified here into a *depth regression network* for inferring the depth map as shown in Fig. 7.12. Regarding the structure of each neural network, the number of convolutional layers and the parameters are exactly the same as Chen's work. Therefore, the reader is referred to [15,51] for a detailed overview. The only difference that I bring to this work concerns the lighting directions that are not included in the inputs of the feature extractor, see Fig. 7.1. I wish to simultaneously achieve non-calibrated PS, as was attempted in [15]. This approach is probably not the most robust [51], however, as a first trial, it can give an idea of the performance that the network can achieve. Unfortunately, the training of the adapted PS-FCN for depth estimation has not been achieved here as the main contribution is based on the dataset and this work is an introduction to future work in this project. Nonetheless, I wish to explain in the next session how the CNN can support a different number of input images for training and testing and then give a quick overview of the two network structures, the loss function and the optimiser used to minimise the error on the predicted depth map.

7.4.1 Shared-Weight Features Extractor

As explained in the deep learning section, the first layers of the neural network will compute a simple function so that the deep layers can compute a complex function. In the shared-weight feature extractor, the principle is the same, the neural network will first extract simple features and then complex ones. To understand what is learned in this specific neural network, G. Chen [17] investigates the features learned to then understand how lighting directions can be derived from the neural network outputs. In that paper, the authors reference the generalised bas-relief (GBR) ambiguity [42] that is inherent to uncalibrated PS imaging. Because usual GBR transformations do not preserve specularities, which are necessary for ambiguity-free lighting estimation,

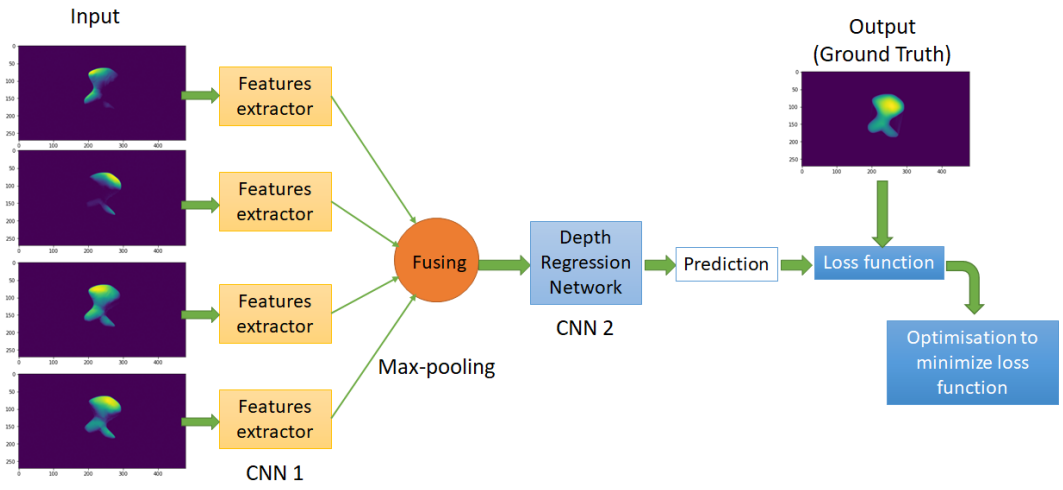


Figure 7.12: Depth estimation Convolutional Neural Network. Four input images, per one depth output, are first fed to the feature extractor. The extracted features from the input images are applied to the fusion layer which consist of a max-pooling that aggregates the features. Finally, the depth regression network predicts the depth map which is then compared to the ground truth using a cosine loss function. This loss function is then optimised to improve the learning.

it is useful to implement a learning-based method that will learn the relation between the specular highlights and the light directions through an end-to-end learning [17]. Fig. 7.13 shows 3 features out of 256 detected by the shared-weight feature extractor which are the specular component (highlight), the attached shadows and the shading [17]. The feature map can be interpreted as a decomposition of the images under different light directions [51] and each feature can provide strong clues for resolving the GBR ambiguity. By explicitly leveraging the object shape and shading information, the estimation of the light direction and intensity can therefore be improved and lead to a self-calibrated neural network where the lighting directions are not needed. For more information on the feature extractor network, the reader can refer to [17].

7.4.2 Max-pooling as a fusion layer

Comparing to [175], G. Chen's CNN can handle a variable number of inputs during training and testing which is not a straightforward task as convolutional layers require the input to have a fixed number of channels during training and testing [15]. Once

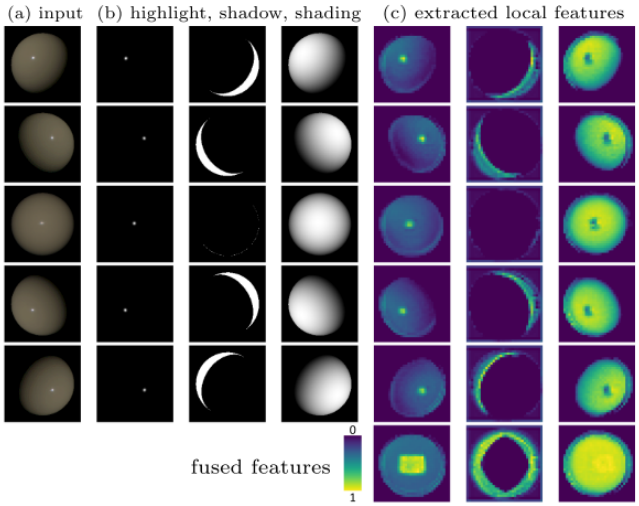


Figure 7.13: Feature visualisation of the *shared-weight feature extractor* on a non-Lambertian sphere. Column 1: 5 of the 96 input images; Columns 2-4: Specular highlight centres, attached shadows, and shading rendered from ground truth; Columns 5-7: 3 of the neural network's 256 feature maps. The last row shows the global features produced by fusing local features with max-pooling. All features are normalised to [0,1] and colour coded [17].

the features are extracted, a fusion layer is needed to aggregate it. The pooling layer mechanism, also called order-agnostic operations, have been used in CNNs to aggregate multi-image information [18]. As explained in [15] and shown in Fig. 7.14, max-pooling operation can extract the most notable information from all the features, while average-pooling can smooth out the prominent and non-activated features. For PS, max-pooling seems to be a better choice as it can naturally aggregate strong features from images captured under different light directions, as highlighted in Fig. 7.13. In addition, an advantage of max-pooling is that it can ignore the non-activated features during training, making it more robust to images incorporating shadow effects. The last row of Fig. 7.13 shows the result of the fusion of 3 local features using max-pooling.

7.4.3 Network architecture

For each input, there is a 3-channel input image with the dimensions of $3 \times h \times w$, where h and w are the image height (270 pixels) and width (480 pixels), respectively. The images are concatenated under the four different light directions, hence by putting

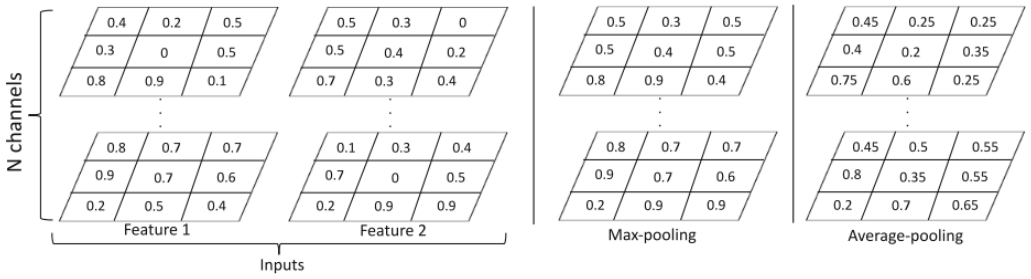


Figure 7.14: Example of max-pooling and average-pooling mechanisms on multi-feature fusions [15].

the images together a $4 \times 3 \times h \times w$ dimensional input to the model is obtained.

Similar to [51], the shared-weight feature extractor has seven convolutional layers, where the feature map is down-sampled twice and then up-sampled once, resulting in a down-sample factor of two. As explained by Chen [51], this design can increase the receptive field and preserve spatial information with a small memory consumption. The normal regression sub-network has four convolutional layers and up-samples the fused feature map to the same spatial dimension as the input images, see Fig. 7.1. An L2-normalisation layer is appended at the end of the depth regression sub-network, just as the normal regression network used in [51], to produce the depth map. A great advantage of the PS-FCN is the adaptability to applied datasets with different image sizes, which makes it fully convolutional.

7.4.4 Loss function and optimiser

Differently from [51], the CNN is supervised by the estimation error between the predicted and the ground truth depth map. The same loss function as [51], called cosine similarity loss, is used as it is a common function in convolutional neural network in general. The loss function equation applied in the introductory work is as follows:

$$L_{depth} = \frac{1}{hw} \sum_{i=1}^{hw} (1 - d_i^T \tilde{d}_i) \tag{7.1}$$

where d_i and $\tilde{d}_i$ correspond to the predicted and the ground truth depth, respectively, at pixel i. If the predicted depth is similar to the ground truth then the dot product

will be close to 1 and the loss will decrease.

Although the training could not be run because of debugging issues and a lack of time, the model is set to use a batch size of 4 for 30 epochs. The commonly used Adam optimiser [176] is set to default setting ($\beta_1 = 0.9$ and $\beta_2 = 0.999$), where the learning rate is initially set to 0.001 and divided by 2 every 5 epochs.

7.5 Future direction on the project

As mentioned earlier, the main contribution in this chapter is the bespoke PS imaging dataset which is a powerful dataset that opens up opportunities to train powerful deep-learning network in the future. Such a dataset has not being reported in the literature and offers the possibility to take a step up in the PS problem where the topography of an object could be retrieved in a second. Work on the CNN implementation is still on going and has not resulted into a full-training of the CNN to display first results. The main challenge encountered here was the dataset generation. This task was demanding in terms of time and organisation. Then, the deep-learning field is itself challenging as it is broad and some skills and knowledge are necessary in different aspects such as the platforms (e.g Keras/Tensorflow or here Pytorch), some basics in gradient descent approach, optimisers, parameters tuning, and so on. The other challenges were the difference in Pytorch between a Linux^{TM} platform and a Windows^{TM} platform and the need of a better GPU to run the code in Python. Most of the code from Chen [51] has been adapted to the work achieved in this thesis, but further debugging need to be finalised to run it on a Windows platform. The work that needs to be carried on to finish improving the computational time of the PS framework can fit in a new PhD project. Here are some guidelines on how to improve it.

7.5.1 Issues on the Windows platform and GPU

A first thing to focus on when continuing this project is to globally check the compatibility between the Linux and Windows platforms. Adapting a code that is commonly run on a Linux platform is not as straightforward as one would think. During the

debugging stage, some errors highlight issues regarding the way to handle workers, for example. Workers are important to separate the data loading process into parallel 'jobs'. When only one worker is in use then all the images are concatenated on the same 'job' which creates a memory issue. However, the worker's behaviour is different between Linux and Windows which creates errors. It is clear that this example is a fixable issue. Some time should be dedicated to the study of Pytorch itself as it is a powerful Python library that can help a developer to put a CNN together faster.

Another important aspect of the project is the GPU selected to run the training of the neural network, as this is crucial when training a CNN. In a CNN, each pixel of each image is read as a model parameter, for example the dataset for the model represents 1,555,200 parameters which requires a big memory. Moreover, after compiling the entire CNN model, this value increases to 2,209,536 model parameters. This number gives an idea of why a powerful GPU is a necessity in such a project and why workers are important to run the learning faster. In Chen's work [51], they use a single NVIDIA Titan X Pascal GPU which, in comparison of the GTX 1060 that could have been available here, is two times more powerful. In addition, the desk computer specifics are as equally important as it needs a good system power supply and a compatible motherboard to operate the GPU properly.

7.5.2 Dataset enhancement

At the moment, the generated dataset only contains objects with low spatial frequencies and soft contours. For the CNN to be robust with real life objects, higher spatial frequencies need to be rendered. Another dataset with OBJ files is made available in the SilNet paper [18] and is called a sculpture dataset. By using the same rendering code on Blender another dataset can be easily generated. Fig. 7.15 shows a selection of four objects from the sculpture dataset and sharper features can be observed which makes it interesting to use in the neural network, especially for shadowing issues to see if the network can be robust enough to deal with this via the max-pooling layer.

By comparing the size of the current dataset (62,904 images rendered) with the one from Chen [51] which represents 5,453,568 rendered images (64 images per blobby and

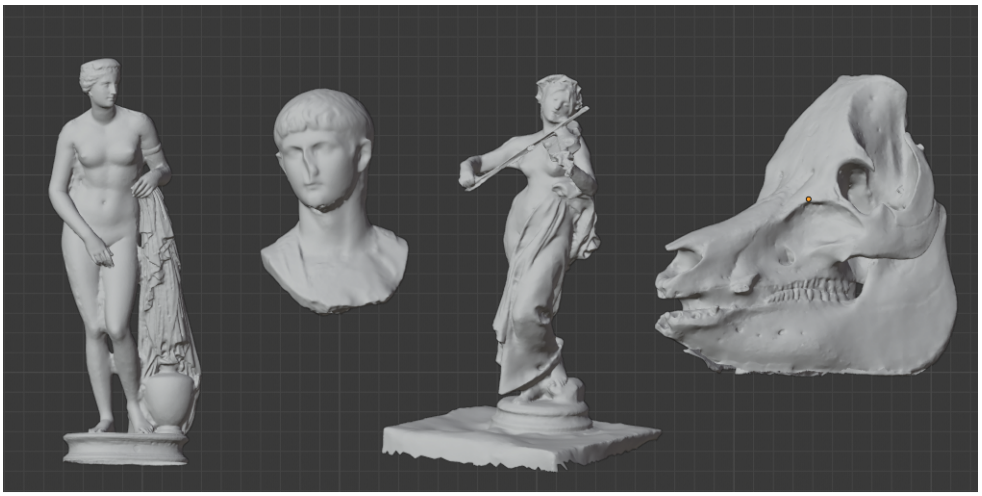


Figure 7.15: SilNet Sculpture dataset [18] selection namely, from left to right, Aphroba statue, Germanicus bust, violin girl statue and okapi skull.

per sculpture sample, which respectively corresponds to 25,920 and 59,292 samples), there is a clear need to increase the size of the dataset. An easy way to create more samples with the blobby dataset is to simply rotate the imported object in Blender; the shapes are not symmetric and new shapes can therefore be rendered. By rotating each object by 45 degrees, the blobby samples can be multiplied by four. The same idea could apply to the sculpture dataset. Again, this needs a good GPU to run fast.

Data augmentation is a strategy commonly used in deep learning. It consists of an image processing procedure to create more images out of already rendered images. Data augmentation can be done through rescaling of images, the mirroring effect or noise perturbation. None of these approaches have been tried here yet, but if the dataset needs to be larger then this would also be a method to use.

7.5.3 Discussion

To date, the main goal of the CNN is to estimate the depth of white Lambertian objects from four images perspective without knowing the lighting directions, in other words non-calibrated photometric stereo assuming Lambertian reflection. As previously mentioned, a hot topic in PS imaging is the reconstruction of non-Lambertian surfaces as most objects in the real-world are not Lambertian. Some work has been carried on

to solve the problem of specular reflection and outliers that makes the estimation of the reflectance map and the surface normal difficult due to the non-linear properties. Some methods do not rely on a neural network and are based on a decomposition of diffuse and specular components of the image [50] or on a compensation strategy [48], for example. However, in the same deep photometric work, Chen [51] achieves surface normal estimation for non-Lambertian surfaces. Now this approach could potentially be implemented as a next step in the fast 3D imaging recovery work.

As explained in [51], a bidirectional reflectance distribution function (BRDF) is a general form for describing the reflectance property of a surface. The MERL dataset [177] contains 100 different BRDF of real-world material and is used in [51] to define a diverse set of surface materials for rendering the blobby and sculpture shapes. A similar procedure could be added to the dataset generation run on Blender. Blender gives the possibility to modify the texture of the object by changing the roughness, the specular and metallic reflection, the transmission of the light and more, so it would be possible to create a wide range of materials to be applied to the imported object and to render four images for each material. A random selection of material would make the dataset evenly distributed.

Another point to consider, after the CNN has been trained a few times, is whether the model is over fitting because of the constant lighting directions. The top-down illumination scenario is important to improve the current experimental setup by having a fast recovery of the depth map. However, to have a deployable system, the CNN model might need to be reviewed to run on different lighting directions.

7.6 Conclusion

The main contribution in this chapter is the generation of a bespoke dataset for PS imaging, which is available in the thesis repository [19]. Moreover, the introduction on deep-learning brings some insight on the potential work that can be carried on to speed up the reconstruction time of the current PS imaging work. A lot of work still needs to be done to make the imaging processing fast and so to achieve real-time 3D imaging. Between non-calibrated PS, non-Lambertian surfaces and making a deployable system,

interesting challenges await the next person that will continue this deep learning work. Deep learning is a powerful tool when used properly, and can potentially achieve in one model several PS challenge, which makes it even more interesting and exciting to use.

Chapter 8

Conclusion and future directions

The work in this thesis demonstrated a proof-of-concept showing that a deployable 3D imaging system can be retrofitted to a room using general lighting and a single camera to achieve high-resolution depth reconstruction of a scene. The fast modulation property of LEDs and the smart Manchester encoding modulation scheme made possible the implementation of a photometric stereo top-down illumination system that does not require any synchronisation between the LED sources itself, nor with the camera. Issues and solutions have been raised and solved on this imaging system to make it more robust to discontinuities via the use of a time-of-flight setup. An attempt to increase the computational time of the 3D imaging process has been reported with the work on deep-learning by first generating a dataset that would initiate the work in this field.

Chapter 1 introduced the 3D imaging state-of-the-art and emphasised the photometric stereo imaging method, with the time-of-flight method and the sensor fusion. In addition the overview on the different sensors, CCD and CMOS, have shown that a careful needs to be made for scientific application when non-linear response or shutter operation can become an issue. In this work, two different cameras have been used, namely the Galaxy S9 and the Fastcam mini camera from Photron. Both sensors rely on a CMOS chip with the possibility to keep a linear operation for the Photron camera. Nonetheless, the 3D PS imaging setup worked perfectly in both cases which means any off-the-shelf can potentially be used. However, the camera needs to reach frame rate of at least 960 fps.

Chapter 2 gave an overview on the LED physics and optical properties and highlighted why these light sources were the best fit for the deployable 3D imaging system. An explanation on the SPAD array also explained the timing capabilities of the chip and how this would be important for a ToF setup.

Chapter 3 presented the mathematical background of photometric stereo imaging along with a demonstration of the numerical method called Fast Marching to integrate surface normal vectors. In this work, assumptions of a Lambertian reflectance and calibrated photometric stereo imaging were made. A new form of self-synchronising modulation called Manchester-Encoded Binary Frequency Division Multiple Access was developed in this theoretical chapter. An explanation on how the modulation scheme is self-clocking was introduced along with how the orthogonality of the scheme allowed each LED to have a specific finger print which made it easy to decode in the demodulation process. An explanation on how this scheme was integrated to the global 3D imaging process was also given.

The experimental synchronization-free top-down illumination photometric stereo imaging proof-of-concept was developed in chapter 4. Results showed that a 2.5D reconstruction of static objects was achieved with an NRMSE error ranging from 1.36% to 19% depending on the shape complexity. Moreover, the LEDs modulation frequency proved to be above visual flicker recognition level which made this approach eye-safe to use in public environment. Thanks to the synchronisation-free feature of the Manchester encoding modulation scheme, the imaging system only requires simple installation of the LEDs on the ceiling, which already exist in general building infrastructures. In addition, because of the use of LEDs, the 3D imaging method was cost-effective and can be hand-held when using the smartphone, which demonstrated the overall ready-to-use aspect.

To complement the work done in chapter 4, chapter 5 demonstrated the dynamic aspect of the top-down illumination PS imaging by using a high-speed camera to record an ellipsoid in motion under the same experimental setup. The idea of this chapter was to achieve real-time imaging. However due to the computation time of the PS process, which was about 3 to 4 minutes, the complete 3D reconstruction system did not result

in it being real-time, but an effective frame rate of 25 fps was achieved. Nonetheless, the PS imaging process and the LED modulation scheme showed to be robust to objects in motion as 3D reconstructed results showed sharp boundaries and well-defined edges shape with an error ranging from 8.8% to 18%. In addition, no synchronisation process was required between the object in motion and the camera as the self-clocking property of the Manchester encoded scheme showed to be very useful in this scenario. However, it was discussed that some alternative can be used, such as structure-from-motion, to improve the frame selection.

In chapter 6, the work focused on solving the discontinuity issues raised in both chapter 4 and 3. A hybrid solution was presented where PS imaging and ToF were merged together to achieve a high-resolution of pre-selected objects by using the ToF as a masking tool to overcome the discontinuous issues of the PS method. A SPAD camera was used to perform time-correlated single photon counting LED-based ToF, which resulted in a cm scale depth map. Without the need to modify the PS pipeline process, this ToF depth map was used as a masking tool by selecting the distance of the object to reconstruct in high-resolution with the PS imaging setup. The root mean square error ranged between 4% and 5% for an object auto-selected from a scene imaged at a distance of 50 cm to 70 cm.

Finally, chapter 7 introduced the idea of using convolutional neural network to improve the reconstruction pipeline of the 3D imaging setup. The main contribution here was the generation of a bespoke dataset for top-down illumination PS imaging. A total of 62,904 images have been rendered with Blender which represent 15,726 singular shapes. The work carried on in this chapter also raised ideas for work to be undertaken in the future, such as uncalibrated or non-Lambertian PS imaging.

8.1 Future work

The work in the thesis is still at a proof-of-concept stage where different methods have been tested to overcome issues such as discontinuities. By coming up with solutions for these problems, equipment that the imaging system would require was added and which does not help with the idea of 'one system fits all' solution. There are therefore

different work paths that need to be investigated to make this 3D imaging system closer to a prototype that could be tested in a shopping centre or a train station, for example. It is clear that the system has only been tried in a lab with specific size 3D printed objects of a specific surface reflectance. These steps were required to show that the system can work under certain assumptions. Now, a list with several ideas that would improve the current synchronisation-free PS imaging system will be listed.

First of all, an easy task would be to combine the colour camera, or smartphone, with the depth camera, or SPAD sensor, to create one imaging device that would only need to be calibrated once. The work done in chapter 5 has shown that the calibration of the two sensors is fast and easy and for the system to be deployable, it will be easier to have one device for all. An infra-red LED could also be integrated to the imaging device to send short-pulses on the scene and be used for the determination of the depth map.

Then, in the Chapter 7, many ideas have been given in the discussion to direct a follow-up project. It is important to say that the deep-learning work would be necessary if the 3D imaging system were to be tested in a real scenario where real-time reconstruction would be needed. Again, the thesis only shows a proof-of-concept, where several assumptions have been made to simplify the imaging system and the determination of the surface normal vectors in the PS imaging process. Real-world applications, such as video surveillance of a train station, would require more advanced systems that can deal with outliers like specular reflections or shadows. Machine learning can be a solution that can include several features to be learned simultaneously. As detailed in the previous section, a convolutional neural network can tackle uncalibrated and non-Lambertian PS imaging at the same time. The choice of surface materials and light source positions would determine how well the neural network would generalise to different scenarios. Some work and understanding on material reflectance, specular reflection, and casting of shadows would be of high value in the future.

Computer vision is a very complex research area and this PhD has demonstrated its broadness by focusing on a very small, yet important, feature which was the deployability of the imaging system by the synchronisation-free aspect between the light

sources and the camera. Another idea to investigate can be the self-positioning feature, enabled by an array of micro-LEDs, between the imaging system and the scene. In other words, having the capability of localising people or objects within the scene to then determine the set of camera/light sources that would be at the best location to image and reconstruct the topography of the target. If the convolutional neural network training were to be a success in achieving uncalibrated PS imaging, then the determination of the light sources' position regarding the target to image would not be required. However, a prototype that would make use of LED arrays to achieve self-positioning of objects or cameras within a room [138] could be envisioned. In addition, by referring to the Fig. 4.1, a smart application of the general lighting could be used for the 3D imaging to enhance a video surveillance system along with a self-positioning structure that would determine which camera to use with which light sources and, to finish, a communication link could also be set within the Manchester Encoding light source or by adding an additional optical link at another wavelength.

This last idea would bring another level of difficulty in the modulation scheme of the light. However the progress made in LiFi [97] does open possibilities on the many ways that the light can be modulated and used for different purposes at once. This multi-purpose imaging/communication optical link can then raise some problems such as the question of security. Once the 3D imaging system is ready for deployment then the matter of the use of the data generated will need to be carefully considered as well.

Bibliography

[1] S. Herbort and C. Wöhler, "An introduction to image-based 3D surface reconstruction and a survey of photometric stereo methods," *3D Research*, vol. 2, no. 3, pp. 1–17, 2011.

[2] J. L. Schonberger and J.-M. Frahm, "Semi-calibrated near-light photometric stereo," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[3] D. Vlasic, P. Peers, I. Baran, P. Debevec, J. Popović, S. Rusinkiewicz, and W. Matusik, "Dynamic shape capture using multi-view photometric stereo," *ACM Transactions on Graphics*, vol. 28, no. 5, p. 1, 2009.

[4] Y. Zhang, G. M. Gibson, R. Hay, R. W. Bowman, M. J. Padgett, and M. P. Edgar, "A fast 3D reconstruction system with a low-cost camera accessory," *Scientific Reports*, vol. 5, pp. 1–7, 2015.

[5] E. Salvador-Balaguer, P. Latorre-Carmona, C. Chabert, F. Pla, J. Lancis, and E. Tajahuerce, "Low-cost single-pixel 3D imaging by using an LED array," *Optics Express*, vol. 26, no. 12, p. 15623, 2018.

[6] W. Sensing AIoT. (2020) Time of Flight: Introduction of direct ToF. [Online]. Available: https://4sense.medium.com/time-of-flight-tof-introduction-of-direct-tof-11ab4c1115ab

[7] M. Peterson. (2020) Iphone 12 pro lidar sensor allows for 6x faster low-light autofocus, instant

ar. [Online]. Available: https://appleinsider.com/articles/20/10/13/iphone-12-pro-lidar-sensor-allows-for-6x-faster-low-light-autofocus-instant-ar

[8] S. Stein. (2021) Lidar is one of the iphone and ipad's coolest tricks. here's what else it can do. [Online]. Available: https://www.cnet.com/tech/mobile/lidar-is-one-of-the-iphone-ipad-coolest-tricks-its-only-getting-better/

[9] S. K. Kim, B. Kang, J. Heo, S.-w. Jung, and O. Choi, "Photometric stereo-based single time-of-flight camera," *Optics Letters*, vol. 39, no. 1, pp. 166–169, 2014.

[10] L. Keal. (2021) Achieving a true global shutter with large format, back illuminated CMOS. [Online]. Available: https://www.princetoninstruments.com

[11] A. D. Griffiths, "Novel optical communications and imaging enabled by cmos interfaced led technology." Ph.D. dissertation, University of Strathclyde, May 2018.

[12] D. Durini, U. Paschen, A. Schwinger, and A. Spickermann, *Silicon based single-photon avalanche diode (SPAD) technology for low-light and high-speed applications.* Elsevier Ltd, 2016.

[13] Y. Quéau, J. D. Durou, and J. F. Aujol, "Variational Methods for Normal Integration," *Journal of Mathematical Imaging and Vision*, vol. 60, no. 4, pp. 609–632, 2018.

[14] W. Basler the power of sight. (2021) Color 3D Point Clouds with Basler blaze. [Online]. Available: https://www.baslerweb.com/en/products/cameras/3d-cameras/blaze-rgb-d/

[15] G. Chen, K. Han, and K.-Y. K. Wong, "PS-FCN: A flexible learning framework for photometric stereo," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–18.

[16] I. Goodfellow, *Deep learning*, ser. Adaptive computation and machine learning, 2016.

[17] G. Chen, M. Waechter, B. Shi, K. Y. K. Wong, and Y. Matsushita, "What Is Learned in Deep Uncalibrated Photometric Stereo?" vol. 12359 LNCS, pp. 745–762, 2020.

[18] A. Zisserman and O. Wiles, "SilNet: Single-and multi-View reconstruction by learning from silhouettes," *The British Machine Vision Association and Society for Pattern Recognition*, 2017.

[19] E. L. Francois. (2022) Thesis code repository. [Online]. Available: https://strathcloud.sharefile.eu/d-s402160cf4fd14b5fa0adb165fcc08dc2

[20] C. Wöhler, *3D Computer Vision*, 2009.

[21] B. Cyganek and J. P. Siebert, *An Introduction to 3D Computer Vision Techniques and Algorithms*, 1st ed. Somerset: Wiley, 2008.

[22] Y. Liu, N. Pears, P. L. Rosin, and P. Huber, *3D imaging, analysis and applications*, 2014, vol. 9781447140.

[23] A. M. Pawlikowska, A. Halimi, R. A. Lamb, and G. S. Buller, "Single-photon three-dimensional imaging at up to 10 kilometers range," *Optics Express*, vol. 25, no. 10, p. 11919, 2017.

[24] A. McCarthy, N. J. Krichel, N. R. Gemmell, X. Ren, M. G. Tanner, S. N. Dorenbos, V. Zwiller, R. H. Hadfield, and G. S. Buller, "Kilometer-range, high resolution depth imaging via 1560 nm wavelength single-photon detection," *Optics Express*, vol. 21, no. 7, p. 8904, 2013.

[25] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2003.

[26] M. Gupta, A. Agrawal, A. Veeraraghavan, and S. G. Narasimhan, "Structured light 3D scanning in the presence of global illumination," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 713–720, 2011.

[27] M. Edgar, S. Johnson, D. Phillips, and M. Padgett, "Real-time computational photon-counting LiDAR," *Optical Engineering*, vol. 57, no. 03, p. 1, 2017.

[28] A. Lipnickas and A. Knys, "A stereovision system for 3-D perception," *Elektronika ir Elektrotechnika*, no. 3, pp. 99–102, 2009.

[29] L. Iocchi and K. Konolige, "A multiresolution stereo vision system for mobile robots," 1998.

[30] S. E. Ghobadi, "Real time object recognition and tracking using 2d/3d images," Ph.D. dissertation, University of Siegen, 2010.

[31] Y. Ji, Y. Li, X. Sun, S. Yan, and N. Guo, "Stereo matching algorithm based on binocular vision," *Proceedings - 2020 7th International Forum on Electrical Engineering and Automation, IFEEA 2020*, pp. 843–847, 2020.

[32] Z. F. Wang and Z. G. Zheng, "A region based stereo matching algorithm using cooperative optimization," *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2008.

[33] K. Law, G. West, P. Murray, and C. Lynch, "3D advanced gas-cooled nuclear reactor fuel channel reconstruction using structure-from-motion," in *10th International Topical Meeting on Nuclear Plant Instrumentation, Control, and Human-Machine Interface Technologies, NPIC and HMIT 2017, NPIC and HMIT 2017*, 2017.

[34] H. Cantzler, "An overview of shape-from-motion," *School of Informatics University of Edinburgh*, 2003.

[35] M. J. Westoby, J. Brasington, N. F. Glasser, M. J. Hambrey, and J. M. Reynolds, "'Structure-from-Motion' photogrammetry: A low-cost, effective tool for geoscience applications," *Geomorphology*, vol. 179, pp. 300–314, 2012.

[36] S. M. Haque, A. Chatterjee, and V. M. Govindu, "High quality photometric reconstruction using a depth camera," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2283–2290, 2014.

[37] J. Ackermann and M. Goesele, "A survey of photometric stereo techniques," *Foundations and Trends in Computer Graphics and Vision*, vol. 9, no. 3-4, pp. 149–254, 2013.

[38] Z. Lu, Y. W. Tai, F. Deng, M. Ben-Ezra, and M. S. Brown, "A 3D imaging framework based on high-resolution photometric-stereo and low-resolution depth," *International Journal of Computer Vision*, vol. 102, no. 1-3, pp. 18–32, 2013.

[39] R. J. Woodham, "Photometric Method For Determining Surface Orientation From Multiple Images," *Optical Engineering*, vol. 19, no. 1, pp. 139–144, 1980.

[40] Y. Matsushita, "Semi-calibrated Photometric Stereo Photometric Stereo," vol. 42, no. 1, p. 7, 2016.

[41] Y. Quéau, B. Durix, T. Wu, D. Cremers, F. Lauze, and J. D. Durou, "LED-Based Photometric Stereo: Modeling, Calibration and Numerical Solution," *Journal of Mathematical Imaging and Vision*, vol. 60, no. 3, 2018.

[42] P. N. Belhumeur, "The Bas-Relief Ambiguity," no. 203, pp. 1–19, 1997.

[43] Y. Quéau, T. Wu, and D. Cremers, "Semi-calibrated near-light photometric stereo," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10302 LNCS, 2017.

[44] F. Logothetis, R. Mecca, Y. Queau, and R. Cipolla, "Near-Field Photometric Stereo in Ambient Light," vol. 2016, no. September, pp. 61.1–61.12, 2017.

[45] P. Favaro and T. Papadhimitri, "A closed-form solution to uncalibrated photometric stereo via diffuse maxima," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 821–828, 2012.

[46] T. Papadhimitri and P. Favaro, "Uncalibrated Near-Light Photometric Stereo," pp. 128.1–128.12, 2015.

[47] S. Sengupta, H. Zhou, W. Forkel, R. Basri, T. Goldstein, and D. Jacobs, "Solving Uncalibrated Photometric Stereo Using Fewer Images by Jointly Optimizing Low-rank Matrix Completion and Integrability," *Journal of Mathematical Imaging and Vision*, vol. 60, no. 4, pp. 563–575, 2018.

[48] Q. Zheng, A. Kumar, B. Shi, and G. Pan, "Numerical reflectance compensation for non-lambertian photometric stereo," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3177–3191, 2019.

[49] K. H. Cheng and A. Kumar, "Revisiting outlier rejection approach for non-lambertian photometric stereo," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1544–1555, 2019.

[50] M. Li, C. yu Diao, D. qing Xu, W. Xing, and D. ming Lu, "A non-Lambertian photometric stereo under perspective projection," *Frontiers of Information Technology and Electronic Engineering*, vol. 21, no. 8, pp. 1191–1205, 2020.

[51] G. Chen, K. Han, B. Shi, Y. Matsushita, and K.-Y. K. Wong, "Deep Photometric Stereo for Non-Lambertian Surfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8828, no. c, 2020.

[52] B. Shi, Z. Mo, Z. Wu, D. Duan, S.-K. Yeung Member, P. Tan, and S. Member, "A Benchmark Dataset and Evaluation for Non-Lambertian and Uncalibrated Photometric Stereo," pp. 1–14, 2018.

[53] G. Schindler, "Photometric Stereo via Computer Screen Lighting for Real-time Surface Reconstruction," *International Symposium on 3D Data Processing, Visualization and Transmission*, pp. 1–6, 2008.

[54] J. Herrnsdorf, L. Broadbent, G. C. Wright, M. D. Dawson, and M. J. Strain, "Video-Rate Photometric Stereo-Imaging with General Lighting Luminaires," *IEEE Photonics Conference*, pp. 483–484, 2017.

[55] V. Argyriou and M. Petrou, "Recursive photometric stereo when multiple shadows and highlights are present," *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, no. 1, 2008.

[56] A. Tontini, L. Gasparini, and M. Perenzoni, "Numerical model of spad-based direct time-of-flight flash lidar CMOS image sensors," *Sensors (Switzerland)*, vol. 20, no. 18, pp. 1–19, 2020.

[57] I. Eichhardt, D. Chetverikov, and Z. Jankó, "Image-guided ToF depth upsampling: a survey," *Machine Vision and Applications*, vol. 28, no. 3-4, pp. 267–282, 2017.

[58] J. Tachella, Y. Altmann, X. Ren, A. McCarthy, G. S. Buller, S. McLaughlin, and J. Y. Tourneret, "Bayesian 3D reconstruction of complex scenes from single-photon lidar data," *SIAM Journal on Imaging Sciences*, vol. 12, no. 1, pp. 521–550, 2019.

[59] M.-C. Amann, T. M. Bosch, M. Lescure, R. A. Myllylae, and M. Rioux, "Laser ranging: a critical review of unusual techniques for distance measurement," *Optical Engineering*, vol. 40, no. 1, pp. 10 – 19, 2001.

[60] C. Mallet and F. Bretar, "Full-waveform topographic lidar: State-of-the-art," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 64, no. 1, pp. 1–16, 2009.

[61] M. J. Sun, M. P. Edgar, G. M. Gibson, B. Sun, N. Radwell, R. Lamb, and M. J. Padgett, "Single-pixel three-dimensional imaging with time-based depth resolution," *Nature Communications*, vol. 7, no. May, pp. 1–6, 2016.

[62] A. D. Griffiths, H. Chen, D. D.-u. Li, R. K. Henderson, J. Herrnsdorf, M. D. Dawson, and M. J. Strain, "Multispectral time-of-flight imaging using light-emitting diodes," *Optics Express*, vol. 27, no. 24, pp. 35 485–35 498, 2019.

[63] S. Foix, G. Alenyà, and C. Torras, "Lock-in time-of-flight (ToF) cameras: A survey," *IEEE Sensors Journal*, vol. 11, no. 9, pp. 1917–1926, 2011.

[64] F. Chiabrando, R. Chiabrando, D. Piatti, and F. Rinaudo, "Sensors for 3D imaging: Metric evaluation and calibration of a CCD/CMOS time-of-flight camera," *Sensors*, vol. 9, no. 12, pp. 10 080–10 096, 2009.

[65] T. Sych, P. Meadows, and B. Allen. (2020) Azure kinect dk depth camera. [Online]. Available: https://docs.microsoft.com/en-us/azure/kinect-dk/depth-camera

[66] Basler. (2022) Basler blaze, time-of-flight camera. [Online]. Available: https://www.baslerweb.com/en/products/cameras/3d-cameras/basler-blaze/

[67] G. Luetzenburg, A. Kroon, and A. A. Bjørk, "Evaluation of the Apple iPhone 12 Pro LiDAR for an Application in Geosciences," *Scientific Reports*, vol. 11, no. 1, pp. 1–9, 2021.

[68] S. Royo and M. Ballesta-Garcia, "An overview of lidar imaging systems for autonomous vehicles," *Applied Sciences (Switzerland)*, vol. 9, no. 19, 2019.

[69] Z.-P. Li, X. Huang, Y. Cao, B. Wang, Y.-H. Li, W. Jin, C. Yu, J. Zhang, Q. Zhang, C.-Z. Peng *et al.*, "Single-photon computational 3d imaging at 45 km," *Photonics Research*, vol. 8, no. 9, pp. 1532–1540, 2020.

[70] R. Koch, A. Kolb, and D. Hutchison, *Time-of-Flight and Depth Imaging: Sensors, Algorithms, and Applications - Dagstuhl 2012 Seminar on Time-of-Flight Imaging and GCPR 2013 Workshop on Imaging New Modalities*, 2013, vol. 8200 LNCS.

[71] L. Li, "Time-of-Flight Camera–An Introduction," *Texas Instruments - Technical White Paper*, no. January, p. 10, 2014.

[72] P. Zanuttigh, G. Marin, C. Dal Mutto, F. Dominio, L. Minto, and G. M. Cortelazzo, *Time-of-flight and structured light depth cameras*. Springer, 2016.

[73] D. B. Lindell, M. O'Toole, and G. Wetzstein, "Single-photon 3D imaging with deep sensor fusion," *ACM Transactions on Graphics*, vol. 37, no. 4, 2018.

[74] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox, "RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments," *International Journal of Robotics Research*, vol. 31, no. 5, pp. 647–663, 2012.

[75] B. Haefner, S. Peng, A. Verma, Y. Queau, and D. Cremers, "Photometric Depth Super-Resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2453–2464, 2019.

[76] A. Bleiweiss and M. Werman, "Fusing time-of-flight depth and color for real-time segmentation and tracking," *Lecture Notes in Computer Science*, vol. 5742 LNCS, pp. 58–69, 2009.

[77] C. Ti, R. Yang, J. Davis, and Z. Pan, "Simultaneous Time-of-Flight sensing and photometric stereo with a single ToF sensor," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June, no. 1, pp. 4334–4342, 2015.

[78] B. Sun, M. P. Edgar, R. Bowman, L. E. Vittert, S. Welsh, A. Bowman, and M. J. Padgett, "3D Computational Imaging with Single-Pixel Detectors," *Science*, vol. 340, no. 6134, pp. 844–847, 2013.

[79] F. Wang and A. Theuwissen, "Linearity analysis of a CMOS image sensor," *IS and T International Symposium on Electronic Imaging Science and Technology*, pp. 84–90, 2017.

[80] B. Tabbert and A. O. Goushcha, "Linearity of the photocurrent response with light intensity for silicon pin photodiode array," in *2006 IEEE Nuclear Science Symposium Conference Record*, vol. 2, 2006, pp. 1060–1063.

[81] Z. Yang, V. Gruev, and J. Van Der Spiegel, "Low fixed pattern noise current-mode imager using velocity saturated readout transistors," *Proceedings - IEEE International Symposium on Circuits and Systems*, no. May, pp. 2842–2845, 2007.

[82] T. Fellers and M. Davidson. (2022) Charge-coupled device linearity. [Online]. Available: https://hamamatsu.magnet.fsu.edu/articles/ccdlinearity.html

[83] A. El Gamal, "High dynamic range image sensors," in *Tutorial at International Solid-State Circuits Conference*, vol. 290, 2002, p. 8.

[84] N. Saha, M. S. Ifthekhar, N. T. Le, and Y. M. Jang, "Survey on optical camera communications: Challenges and opportunities," *IET Optoelectronics*, vol. 9, no. 5, pp. 172–183, 2015.

[85] M. V. RadhaKrishna, M. Venkata Govindh, and P. Krishna Veni, "A review on image processing sensor," *Journal of Physics: Conference Series*, vol. 1714, no. 1, 2021.

[86] L. C. P. Gouveia and B. Choubey, "A review on image processing sensor," *Sensor Review*, vol. 36, no. 3, 2016.

[87] Photron. (2022) Fastcam mini ux. [Online]. Available: https://photron.com/fastcam-mini-ux/

[88] H. J. Round, "A note on carborundum," *Electrical World*, 1907.

[89] The 2014 Nobel Prize in Physics - Press Release. [Online]. Available: https://www.nobelprize.org/nobel_prizes/physics/laureates/2014/press.html

[90] J. Grubor, S. Randel, K.-d. Langer, and J. W. Walewski, "Broadband Information Broadcasting Using," *Journal of Lightwave Technology*, vol. 26, no. 24, pp. 3883–3892, 2008.

[91] J. J. D. McKendry, B. R. Rae, Z. Gong, K. R. Muir, B. Guilhabert, D. Massoubre, E. Gu, D. Renshaw, M. D. Dawson, and R. K. Henderson, "Individually addressable AlInGaN micro-LED arrays with CMOS control and subnanosecond output pulses," *IEEE Photonics Technology Letters*, vol. 21, no. 12, pp. 811–813, 2009.

[92] Z. J. Liu, K. M. Wong, C. W. Keung, C. W. Tang, and K. M. Lau, "Monolithic LED Microdisplay on Active Matrix Substrate Using Flip-Chip Technology," *J. Sel. Topics Quantum Electron.*, vol. 15, no. 4, pp. 1298–1302, 2009.

[93] G.-b. Leds, S. Zhang, S. Watson, J. J. D. Mckendry, D. Massoubre, A. Cogman, E. Gu, R. K. Henderson, A. E. Kelly, and M. D. Dawson, "1.5 Gbit/s Multi-Channel Visible Light Communications Using CMOS-Controlled GaN-Based LEDs," vol. 31, no. 8, pp. 1211–1216, 2013.

[94] A. D. Griffiths, J. Herrnsdorf, C. Lowe, M. Macdonald, R. Henderson, M. J. Strain, and M. D. Dawson, "Poissonian communications: free space optical data transfer at the few-photon level," *arXiv: Optics*, 2018.

[95] J. Herrnsdorf, M. J. Strain, E. Gu, R. K. Henderson, and M. D. Dawson, "Positioning and space-division multiple access enabled by structured illumination with light-emitting diodes," *Journal of Lightwave Technology*, vol. 35, no. 12, pp. 2339–2345, 2017.

[96] J. Herrnsdorf, J. McKendry, M. Stonehouse, L. Broadbent, G. C. Wright, M. D. Dawson, and M. J. Strain, "Led-based photometric stereo-imaging employing frequency-division multiple access," in *2018 IEEE Photonics Conference (IPC)*, Sep. 2018, pp. 1–2.

[97] H. Haas, L. Yin, Y. Wang, and C. Chen, "What is LiFi?" *Journal of Lightwave Technology*, vol. 34, no. 6, pp. 1533–1544, 2016.

[98] M. H. Crawford, "LEDs for solid-state lighting: Performance challenges and recent advances," *IEEE Journal on Selected Topics in Quantum Electronics*, vol. 15, no. 4, pp. 1028–1040, 2009.

[99] J. McKendry, "Micro-pixellated AlInGaN light-emitting diode arrays for optical communications and time-resolved fluorescence lifetime measurements," no. April, p. 199, 2011.

[100] E. F. Schubert, *Light-Emitting Diodes*. Cambridge: Cambridge University Press, 2006.

[101] D. Halliday, *Fundamentals of physics*, extended ed. New York: Wiley, 2018.

[102] M. Boroditsky and E. Yablonovitch, "Light-emitting diode extraction efficiency," *Light-Emitting Diodes: Research, Manufacturing, and Applications*, vol. 3002, no. 2, pp. 119–122, 1997.

[103] J. Herrnsdorf, J. J. McKendry, E. Xie, M. J. Strain, I. M. Watson, E. Gu, and M. D. Dawson, "High speed spatial encoding enabled by CMOS-controlled micro-LED arrays," *2016 IEEE Photonics Society Summer Topical Meeting Series, SUM 2016*, vol. 2, pp. 173–174, 2016.

[104] J. J. D. McKendry, R. P. Green, A. E. Kelly, Z. Gong, B. Guilhabert, D. Massoubre, E. Gu, and M. D. Dawson, "High-speed visible light communications using individual pixels in a micro light-emitting diode array," *IEEE Photonics Technology Letters*, vol. 22, no. 18, pp. 1346–1348, 2010.

[105] J. J. D. McKendry, D. Massoubre, S. Zhang, B. R. Rae, R. P. Green, E. Gu, R. K. Henderson, A. E. Kelly, and M. D. Dawson, "Visible-Light Communications Using a CMOS-Controlled Micro-Light- Emitting-Diode Array," *Journal of Lightwave Technology*, vol. 30, no. 1, pp. 61–67, 2012.

[106] A. D. Griffiths, J. Herrnsdorf, M. J. Strain, and M. D. Dawson, "Scalable visible light communications with a micro-led array projector and high-speed smartphone camera," *Opt. Express*, vol. 27, no. 11, pp. 15 585–15 594, May 2019.

[107] A. D. Griffiths, M. S. Islim, J. Herrnsdorf, J. J. D. McKendry, R. Henderson, H. Haas, E. Gu, and M. D. Dawson, "CMOS-integrated GaN LED array for discrete power level stepping in visible light communications," *Optics Express*, vol. 25, no. 8, p. A338, 2017.

[108] R. K. Henderson, N. Johnston, F. Mattioli, D. Rocca, H. Chen, D. D.-u. Li, G. Hungerford, R. Hirsch, D. Mcloskey, P. Yip, and D. J. S. Birch, "A 192 × 128 Time Correlated SPAD Image Sensor in 40-nm CMOS Technology," *IEEE Journal of Solid-State Circuits*, vol. PP, pp. 1–10, 2019.

[109] "Photomultiplier Tubes, Third Edition," *Hamamatsu Photonics K.K. Electron Tube Division*, 2007.

[110] M. M. Ombaba, H. Karaagac, K. G. Polat, and M. S. Islam, *Nanowire enabled photodetection.* Elsevier Ltd, 2016.

[111] N. Dinu, *Silicon photomultipliers (SiPM).* Elsevier Ltd, 2016.

[112] S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa, "Avalanche photodiodes and quenching circuits for single-photon detection," *Appl. Opt.*, vol. 35, no. 12, pp. 1956–1976, Apr 1996.

[113] S. Cova, A. Longoni, and A. Andreoni, "Towards picosecond resolution with single-photon avalanche diodes," *Review of Scientific Instruments*, vol. 52, no. 3, pp. 408–412, 1981.

[114] Y. Quéau, J. D. Durou, and J. F. Aujol, "Normal Integration: A Survey," *Journal of Mathematical Imaging and Vision*, vol. 60, no. 4, pp. 576–593, 2018.

[115] W. Xie, C. Dai, and C. C. Wang, "Photometric stereo with near point lighting: A solution by mesh deformation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June, pp. 4585–4593, 2015.

[116] J. Ho, J. Lim, M.-H. Yang, and D. Kiriegma, "Integrating surface normal vectors using fast marching method," *Computer Vision-ECCV*, pp. 239–250, 2006.

[117] B. K. Horn and M. J. Brooks, "The variational approach to shape from shading," *Computer Vision, Graphics and Image Processing*, vol. 33, no. 2, pp. 174–208, 1986.

[118] E. Rouy and A. Tourin, "Viscosity solutions approach to shape-from-shading," *SIAM Journal on Numerical Analysis*, vol. 29, no. 3, pp. 867–884, 1992.

[119] Z. Wu and L. Li, "A line-integration based method for depth recovery from surface normals." *Computer Vision, Graphics and Image Processing*, vol. 43, pp. 53–66, 1988.

[120] R. T. Frankot and R. Chellappa, "A Method for Enforcing Integrability in Shape from Shading Algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10, no. 4, pp. 439–451, 1988.

[121] I. Horovitz and N. Kiryati, "Depth from gradient fields and control points: Bias correction in photometric stereo," *Image and Vision Computing*, vol. 22, no. 9, pp. 681–694, 2004.

[122] R. Klette and K. Schluens, "Height data from gradient maps," *Proceedings of SPIE*, vol. 2908, no. 1, pp. 204–215, 1996.

[123] T. Simchony, R. Chellappa, and M. Shao, "Direct Analytical Methods for Solving Poisson Equations in Computer Vision Problems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 5, pp. 435–446, 1990.

[124] J.-D. Durou, J.-F. Aujol, and F. Courteille, "Integration of a Normal Field in the Presence of Discontinuities," *Proceedings of the 7^{th} International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, vol. 5681, pp. 261–273, 2009.

[125] M. Breuss, Y. Quéau, M. Bähr, and J.-D. Durou, "Highly Efficient Surface Normal Integration," *Algoritmy Conference on Scientific Computing (ALGORITMY)*, no. March, pp. 204–213, 2016.

[126] S. Galliani and M. Breuß, "Fast and Robust Surface Normal Integration by a Discrete Eikonal Equation," *British Machine Vision Conference 2012*, no. 6, p. 2009, 2009.

[127] M. Bähr, M. Breuß, Y. Quéau, A. S. Boroujerdi, and J. D. Durou, "Fast and accurate surface normal integration on non-rectangular domains," *Computational Visual Media*, vol. 3, no. 2, pp. 107–129, 2017.

[128] J. A. Sethian, *Level Set Methods and Fast Marching Methods: evolving interfaces in computational geometry, fluid mechanics, computer vision, and material*

science, ser. Cambridge University. Cambridge monographs on applied and computational mathematics, 1999.

[129] *Partial Differential Equations.* Providence, R.I.: American Mathematical Society, 1998.

[130] M. G. Crandall and P.-L. Lions, "Viscosity Solutions of Hamilton-Jacobi Equations," *Transactions of the American Mathematical Society*, vol. 277, no. 1, p. 1, 1983.

[131] C. Hirsch, *Numerical computation of internal and external flows*, ser. Wiley series in numerical methods in engineering. Chichester [England] ; New York: Wiley, 1988.

[132] R. Courant, E. Isaacson, and M. Rees, "On the solution of nonlinear hyperbolic differential equations by finite differences," *Communications on Pure and Applied Mathematics*, vol. 5, no. 3, pp. 243–255, 1952.

[133] R. Sedgewick, *ALGORITHMS.* Addison-Wesley Publishing Company, 1983.

[134] J. A. Sethian, "A fast marching level set method for monotonically advancing fronts." *Proceedings of the National Academy of Sciences*, vol. 93, no. 4, pp. 1591–1595, 1996.

[135] E. Ziegel, W. Press, B. Flannery, S. Teukolsky, and W. Vetterling, *Numerical Recipes: The Art of Scientific Computing.* Cambridge University, 1987.

[136] J. Cardelino. (2018) 2D fast marching algorithm. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/18529-2d-fast-marching-algorithm

[137] J. K. Park, T. G. Woo, M. Kim, and J. T. Kim, "Hadamard Matrix Design for a Low-Cost Indoor Positioning System in Visible Light Communication," *IEEE Photonics Journal*, vol. 9, no. 2, 2017.

[138] J. Herrnsdorf, M. J. Strain, E. Gu, R. K. Henderson, and M. D. Dawson, "Positioning and space-division multiple access enabled by structured illumination with light-emitting diodes," *Journal of Lightwave Technology*, vol. 35, no. 12, pp. 2339–2345, 2017.

[139] C. Hernández, G. Vogiatzis, G. J. Brostow, B. Stenger, and R. Cipolla, "Non-rigid photometric stereo with colored light," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1–8, 2007.

[140] T.-H. Do and M. Yoo, "An in-depth survey of visible light communication based positioning systems," *Sensors*, vol. 16, no. 5, p. 678, 2016.

[141] A. Singh, *Introduction to matrix theory*, 2021.

[142] E. L. Francois, J. Herrnsdorf, J. J. D. McKendry, L. Broadbent, G. Wright, M. D. Dawson, and M. J. Strain, "Synchronization-free top-down illumination photometric stereo imaging using light-emitting diodes and a mobile device," *Opt. Express*, vol. 29, no. 2, pp. 1502–1515, Jan 2021.

[143] Optica. (2021) Researchers Acquire 3D Images with LED Room Lighting and a Smartphone. [Online]. Available: https://www.optica.org/en-us/about/newsroom/news_releases/2021/researchers_acquire_3d_images_with_led_room_lighti/

[144] J. Herrnsdorf, J. McKendry, M. Stonehouse, L. Broadbent, G. C. Wright, M. D. Dawson, and M. J. Strain, "Lighting as a service that provides simultaneous 3d imaging and optical wireless connectivity," in *2018 IEEE Photonics Conference (IPC)*, 2018, pp. 1–2.

[145] M. Chantler and J. Wu, "Rotation Invariant Classification of 3D Surface Texturesusing Photometric Stereo and Surface Magnitude Spectra," *BMVC2000*, pp. 49.1–49.10, 2013.

[146] Blender. (2021) Blender software. [Online]. Available: https://www.blender.org/

[147] B. Kneis and W. Zhang, "3D Face Recognition using Photometric Stereo and Deep Learning," *ACM International Conference Proceeding Series*, vol. Part F1625, pp. 255–261, 2020.

[148] T. Kovacovsky, "Parameters of 3D sensing techniques in a nutshell," no. September, p. 7, 2017.

[149] W. Xie, Z. Song, and R. Chung, "Real-time three-dimensional fingerprint acquisition via a new photometric stereo means," *Optical Engineering*, vol. 52, no. 10, p. 103103, 2013.

[150] "Motion cam 3d," https://www.photoneo.com/motioncam-3d/, accessed: 2021-06-18.

[151] J. Geng, "Structured-light 3D surface imaging: a tutorial," *Advances in Optics and Photonics*, vol. 3, no. 2, p. 128, 2011.

[152] T. Yoda, H. Nagahara, R. I. Taniguchi, K. Kagawa, K. Yasutomi, and S. Kawahito, "The dynamic photometric stereo method using a multi-tap CMOS image sensor," *Sensors (Switzerland)*, vol. 18, no. 3, pp. 1–17, 2018.

[153] J. Tachella, Y. Altmann, N. Mellado, A. Mccarthy, J.-y. Tourneret, S. Mclaughlin, R. Tobin, and G. S. Buller, "Real-time 3D reconstruction from single-photon lidar data using plug-and-play point cloud denoisers," *Nature Communications*, no. 2019, pp. 1–6.

[154] T. Weise, B. Leibe, and L. Van Gool, "Fast 3D scanning with automatic motion compensation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007.

[155] T. Malzbender, B. Wilburn, D. Gelb, and B. Ambrisco, "Surface enhancement using real-time photometric stereo and reflectance transformation," *Eurographics Symposium on Rendering Techniques*, pp. 245–150, 2006.

[156] M. L. Smith and L. N. Smith, "Dynamic photometric stereo - A new technique for moving surface analysis," *Image and Vision Computing*, vol. 23, no. 9, pp. 841–852, 2005.

[157] E. L. Francois. (2020) Dataset 1. [Online]. Available: https://doi.org/10.15129/44b337ab-9fd9-4983-9079-a0dcfaae1d84

[158] R. Cipolla, S. Battiato, and G. M. Farinella, "Computer Vision: Detection, Recognition and Reconstruction," *Springer: Berlin, Germany*, vol. 285, 2010.

[159] C. Zhang, S. Lindner, I. M. Antolovic, M. Wolf, and E. Charbon, "A CMOS SPAD imager with collision detection and 128 dynamically reallocating TDCs for single-photon counting and 3D time-of-flight imaging," *Sensors (Switzerland)*, vol. 18, no. 11, 2018.

[160] K. Morimoto, A. Ardelean, M.-L. Wu, A. C. Ulku, I. M. Antolovic, C. Bruschini, and E. Charbon, "Megapixel time-gated SPAD image sensor for 2D and 3D imaging applications," *Optica*, vol. 7, no. 4, p. 346, 2020.

[161] T. Chan and L. Vese, "Active contours without edges," *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 266–277, 2001.

[162] Q. Zhang, M. Ye, R. Yang, Y. Matsushita, B. Wilburn, and H. Yu, "Edge-preserving photometric stereo via depth fusion," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2472–2479, 2012.

[163] C. Hernández, G. Vogiatzis, and R. Cipolla, "Shadows in Three-Source Photometric Stereo," *Computer Vision – ECCV*, pp. 290–303, 2010.

[164] H. Santo, M. Samejima, Y. Sugano, B. Shi, and Y. Matsushita, "Deep Photometric Stereo Network," *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017*, vol. 2018-January, pp. 501–509, 2017.

[165] G. Chen, K. Han, B. Shi, and Y. M. K.-y. K. Wong, "Self-calibrating Deep Photometric Stereo Networks," *IEEE CVF CVPR*, pp. 8731–8739, 2019.

[166] Y. Tang, R. Salakhutdinov, and G. Hinton, "Deep lambertian networks," *arXiv preprint arXiv:1206.6445*, 2012.

[167] L. Lu, L. Qi, Y. Luo, H. Jiao, and J. Dong, "Three-dimensional reconstruction from single image base on combination of CNN and multi-spectral photometric stereo," *Sensors (Switzerland)*, vol. 18, no. 3, 2018.

[168] F. Logothetis, I. Budvytis, R. Mecca, and R. Cipolla, "A CNN based approach for the near-field photometric stereo problem," *arXiv preprint arXiv:2009.05792*, 2020.

[169] S. Ikehata, "Cnn-ps: Cnn-based photometric stereo for general non-convex surfaces," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–18.

[170] Y. Ju, L. Qi, H. Zhou, J. Dong, and L. Lu, "Demultiplexing colored images for multispectral photometric stereo via deep neural networks," *IEEE Access*, vol. 6, pp. 30 804–30 818, 2018.

[171] C. R. Gonzalez and Y. S. Abu-Mostafa, "Mismatched Training and Test Distributions Can Outperform Matched Ones," *Neural Computation*, vol. 27, pp. 365–387, 2015.

[172] A. Fuchs, C. Knoll, and F. Pernkopf, "Distribution Mismatch Correction for Improved Robustness in Deep Neural Networks," 2021.

[173] M. K. Johnson and E. H. Adelson, "Shape estimation in natural illumination," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2553–2560, 2011.

[174] B. Shi, Z. Mo, Z. Wu, D. Duan, S. K. Yeung, and P. Tan, "A Benchmark Dataset and Evaluation for Non-Lambertian and Uncalibrated Photometric Stereo," *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 271–284, 2019.

[175] D. Eigen, "Depth Map Prediction from a Single Image using a Multi-Scale Deep Network," pp. 1–9.

[176] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *ICLR*, pp. 1–15, 2015.

[177] W. Matusik, H. Pfister, M. Brand, and L. McMillan, "A data-driven reflectance model," *ACM SIGGRAPH 2003 Papers, SIGGRAPH '03*, pp. 759–769, 2003.

Appendix A

Appendix

In this appendix, the illustration of the MEB-FDMA encoding scheme for two emitters is given. In a following session, the proof for the equivalence between Eq. (3.16) and Eq. (3.17) and the proof for Eq. (3.26) are provided.

A.1 Illustration of the example of two emitters

Consider two emitters, $N = 2$. Then according to Eq.(3.21) starting with two simple square waves:

$$s_{i,j}^{(0)} = \begin{bmatrix} -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix} \tag{A.1}$$

Then the Manchester encoded transmitter signal s is, according to Eq.(3.22):

$$s_{i,j} = \begin{bmatrix} 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \end{bmatrix} \tag{A.2}$$

The decoding matrices $D^{(i)}$ are calculated by Eqs. (3.25) and (3.27) to:

$$D^{(1)} = \begin{bmatrix} 0.35 & -0.35 & -0.35 & 0.35 \\ 0.35 & 0.35 & -0.35 & -0.35 \\ 0.35 & -0.35 & -0.35 & 0.35 \\ 0.35 & 0.35 & -0.35 & -0.35 \end{bmatrix} \tag{A.3}$$

$$D^{(2)} = \begin{bmatrix} 0.35 & -0.35 & 0.35 & -0.35 \\ 0.61 & 0.20 & -0.20 & 0.20 \\ 0 & 0.58 & 0.29 & -0.29 \\ 0 & 0 & 0.5 & 0.5 \\ -0.35 & 0.35 & -0.35 & 0.35 \\ -0.61 & -0.20 & 0.20 & -0.20 \\ 0 & -0.58 & -0.29 & 0.29 \\ 0 & 0 & -0.5 & -0.5 \end{bmatrix} \tag{A.4}$$

The numbers in Eqs. (A.3) and (A.4) are not given with full numerical precision.

Let us now look at a specific example, where $I_1 = 1$ and $I_2 = 2$, and emitter 1 has a phase-shift of 0.3 and emitter 2 a phase shift of 1.7. Then the received signal r is calculated by Eqs. (3.23) and (3.24):

$$r = \begin{bmatrix} 0.2 & 1.6 & -0.2 & -0.4 & 1.8 & -2.4 & -1.8 & 1.2 \end{bmatrix} \tag{A.5}$$

Note that an arbitrary DC offset can be added to r without compromising the evaluation below. The orthogonal components of r according to Eq. (3.28) are:

$$c^{(1)} = \begin{bmatrix} 1.98 & 0.85 \end{bmatrix} \tag{A.6}$$

$$c^{(2)} = \begin{bmatrix} -0.85 & -0.82 & 3.23 & 0 \end{bmatrix} \tag{A.7}$$

The contributions of each emitter to r are then calculated by Eq. (3.29):

$$r^{(1)} = \begin{bmatrix} 1 & -0.4 & -1 & 0.4 & 1 & -0.4 & -1 & 0.4 \end{bmatrix} \tag{A.8}$$

$$r^{(2)} = \begin{bmatrix} -0.8 & 2 & 0.8 & -0.8 & 0.8 & -2 & -0.8 & 0.8 \end{bmatrix} \tag{A.9}$$

Then finally correct decoding is done with Eq. (3.30):

$$I_1 = \max(r^{(1)}) = 1 \tag{A.10}$$

$$I_2 = \max(r^{(2)}) = 2 \tag{A.11}$$

A.2 Proofs

A.2.1 Proof of equivalence of Eqs. (3.16) and (3.17)

By setting $\alpha = 0$, (3.16)⇒(3.17) in the main text is directly shown. For the other direction, the left hand side of Eq. (3.16) is reordered:

$$\begin{aligned} &\sum_{j=1}^{n} s_{i,j} \left((1-\alpha) s_{i',1+(j-1+k)\%n} + \alpha s_{i',1+(j+k)\%n} \right) \\ =&(1-\alpha) \sum_{j=1}^{n} s_{i,j} s_{i',1+(j-1+k)\%n} \\ &+ \alpha \sum_{j=1}^{n} s_{i,j} s_{i',1+(j+k)\%n} \end{aligned} \tag{A.12}$$

If Eq. (3.17) is fulfilled, then the right hand side of Eq. (A.12) is 0 and thus (3.16)⇐(3.17) is shown.

A.2.2 Proof of Eq. (3.26)

Lemma 1. *Given a row vector $s \in \mathbb{R}^n$, define $k_0 \in \{1, \ldots, n\}$ via Eq. (A.13):*

$$\begin{aligned} k_0 = \min \{ k \in \{1, \ldots, n\} \,|\, &\exists \beta \in \mathbb{R}, \forall i = 1, \ldots, n : \\ &s_{1+(i-1+k)\%n} = \beta s_i \} \end{aligned} \tag{A.13}$$

Then:

1. $\beta \in \{-1, 1\}$
2. *s is periodic. If $\beta = 1$ then the period is k_0 and if $\beta = -1$ then the period is $2k_0$*
3. *If a matrix S is constructed from s using Eq. (3.25), then its rank is* $\operatorname{rank}(S) \leq k_0$

Proof. The modulus of s is calculated:

$$\begin{aligned} |s|^2 = \sum_{i=0}^{n} s_i^2 = \sum_{i=0}^{n} s_{1+(i-1+k_0)\%n}^2 = \beta^2 |s|^2 \\ \Rightarrow |\beta| = 1 \end{aligned} \tag{A.14}$$

The periodicity of s follows directly from Eq. (A.14).

The periodicity of s directly implies that $\operatorname{rank}(S) \leq k_0$, because all the rows from k_0 on are periodic repeats of the rows 1 to k_0. □

Now let $k = k_r + q2^i, \quad k_r < 2^i, q \in \mathbb{N}$. Then Eq. (3.17) gives:

$$s_{i,j+k} = (-1)^q s_{i,j+k_r} \tag{A.15}$$

The special case of $k_r = 0$ implies directly that in Lemma 1:

$$k_0 \leq 2^i \tag{A.16}$$

and therefore following Lemma 1:

$$\operatorname{rank}(S^{(i)}) \leq 2^i \tag{A.17}$$

Lemma 2. *If a matrix $S^{(i)}$ is constructed from s using Eq. (3.25), and an $2^i \times 2^i$ square matrix $\hat{S}^{(i)}$ is defined by:*

$$\hat{S}_{k,j}^{(i)} = S_{k,j}^{(i)} \quad k = 1, \ldots, 2^i \quad j = 1, \ldots, 2^i \tag{A.18}$$

Then $\hat{S}^{(i)}$ has full rank.

Proof. Consider the equation system:

$$\sum_{k=1}^{2^i} c_k s_{i,j-1+k} = 0 \qquad \forall j = 1, \ldots, 2^i \tag{A.19}$$

$\hat{S}^{(i)}$ has full rank, if and only if Eq. (A.19) implies that all c_k are 0, which is what is going to shown here. Eq. (A.19) can be written for even and odd j, following Eq. (3.17):

$$0 = \sum_{l=1}^{2^{i-1}} \left[c_{2l-1}(-1)^{\lceil \frac{j+2l-2}{2^i} \rceil} + c_{2l}(-1)^{1+\lceil \frac{j+2l}{2^i} \rceil} \right],$$
$$(j = 2p) \tag{A.20}$$
$$0 = \sum_{l=1}^{2^{i-1}} \left[c_{2l-1}(-1)^{1+\lceil \frac{j+2l-1}{2^i} \rceil} + c_{2l}(-1)^{\lceil \frac{j-1+2l}{2^i} \rceil} \right],$$
$$(j = 2q-1) \tag{A.21}$$
$$p = 1, \ldots, 2^{i-1} \tag{A.22}$$
$$q = 1, \ldots, 2^{i-1} \tag{A.23}$$

The right hand sides of Eqs. (A.20) and (A.21) are zero and thus equal to each other, yield the equation set (A.24):

$$\begin{aligned} 0 = &\sum_{l=1}^{2^{i-1}} c_{2l-1} \left[(-1)^{\lceil \frac{p+l-1}{2^{i-1}} \rceil} + (-1)^{\lceil \frac{q+l-1}{2^{i-1}} \rceil} \right] \\ &- \sum_{l=1}^{2^{i-1}} c_{2l} \left[(-1)^{\lceil \frac{p+l}{2^{i-1}} \rceil} + (-1)^{\lceil \frac{q+l-1}{2^{i-1}} \rceil} \right] \end{aligned} \tag{A.24}$$

The c_k can be determined from Eq. (A.24) by iteratively looking at specific values of p and q. The following iteration is performed:

Initial conditions for iteration

Consider $p = 2^{i-1}, q = 2^{i-1}$, then Eq. (A.24) is:

$$0 = -c_1 + \sum_{l=2}^{2^{i-1}} [c_{2l} - c_{2l-1}] \tag{A.25}$$

Appendix A. Appendix

Consider $p = 1, q = 2^{i-1}$, then Eq. (A.24) is:

$$0 = c_1 - c_2 + 2_{2^i} \tag{A.26}$$

Consider $p = 2^{i-1}, q = 1$, then Eq. (A.24) is:

$$0 = c_1 \tag{A.27}$$

Consider $p = 1, q = 1$, then Eq. (A.24) is:

$$0 = \sum_{l=1}^{2^{i-1}-1} [c_{2l-1} - c_{2l}] + c_{2^{i-1}-1} \tag{A.28}$$

Equations (A.25), (A.26), (A.27), and (A.28) together imply:

$$c1 = c2 = c_{2^i} = 0 \tag{A.29}$$

With this knowledge, consider again $p = 1, q = 1$, then Eq. (A.24) is:

$$0 = \sum_{l=2}^{2^{i-1}-1} c_{2l} - \sum_{l=2}^{2^{i-1}} c_{2l-1} \tag{A.30}$$

Consider $p = 1, q = 2$, then Eq. (A.24) is:

$$0 = \sum_{l=2}^{2^{i-1}-1} [c_{2l} - c_{2l-1}] + c_{2^i-1} \tag{A.31}$$

Equations (A.30) and (A.31) imply:

$$c_{2^i-1} = 0 \tag{A.32}$$

Iteration Step m.a

It is known that:

$$c1 = 0, c_2 = 0, c_k = 0 \qquad \forall k > 2^i - 2(m-1) \tag{A.33}$$

Consider $p = m-1, q = m-1$, then Eq. (A.24) is:

$$0 = \sum_{l=2}^{2^{i-1}-m+1} [c_{2l} - c_{2l-1}] \tag{A.34}$$

Consider $p = m, q = m-1$, then Eq. (A.24) is:

$$0 = \sum_{l=2}^{2^{i-1}-m} c_{2l} - \sum_{l=2}^{2^{i-1}-m+1} c_{2l-2m+1} \tag{A.35}$$

Equations (A.34) and (A.35) imply:

$$c_{2^i-2m+2} = 0 \tag{A.36}$$

Iteration Step m.b

Consider $p = m, q = m$, then Eq. (A.24) is:

$$0 = \sum_{l=2}^{2^{i-1}-m} c_{2l} - \sum_{l=2}^{2^{i-1}-m+1} c_{2l-1} \tag{A.37}$$

Consider $p = m, q = m+1$, then Eq. (A.24) is:

$$0 = \sum_{l=2}^{2^{i-1}-m} [c_{2l} - c_{2l-1}] + c_{2^i-2m+1} \tag{A.38}$$

Equations (A.37) and (A.38) imply:

$$c_{2^i-2m+1} = 0 \tag{A.39}$$

(Iteration finished)

By repeating steps m.a and m.b iteratively for $m = 2, \ldots, 2^{i-1}-1$ it is shown that:

$$c_k = 0 \qquad \forall k = 1, \ldots, 2^i \tag{A.40}$$

And thus, $\hat{S}^{(i)}$ must have full rank.

Remark: It has been observed empirically that the determinant of $\hat{S}^{(i)}$ is $\det(\hat{S}^{(i)}) = 2^{2^i-1}$ for $i = 1, \ldots, 8$, but this relationship has not been proven in general. □

Note that $S^{(i)}$ has at least the same rank or higher rank than $\hat{S}^{(i)}$ in Lemma 2 and therefore:

$$\operatorname{rank}(S^{(i)}) \geq \operatorname{rank}(\hat{S}^{(i)}) = 2^i \tag{A.41}$$

Equations (A.17) and (A.41) prove Eq. (3.26).

A.2.3 Phase invariant orthogonality and the Kronecker product

In this section it is proven that if s is phase invariant orthogonal, then $t = s \otimes v$ is also phase invariant orthogonal for any arbitrary row vector v. Let m be the length of v and $k = pm + q$, $p = 0, \ldots, n-1$, $q = 0, \ldots, m-1$ then the following can be calculated:

$$\begin{aligned}
&\sum_{j=1}^{nm} t_{i,j} t_{i',1+(j+k)\%(nm)} \\
=&\sum_{j=1}^{n}\sum_{l=1}^{m} t_{i,m(j-1)+l} t_{i',1+(m(j-1)+l+k)\%(nm)} \\
=&\sum_{l=1}^{m} v_l v_{1+(l+q)\%m} \sum_{j=1}^{n} s_{i,j} s_{i',1+(j+p)\%n} \\
=&0 \qquad \forall i \neq i'
\end{aligned} \tag{A.42}$$

Eq. (A.42) has been used in Eq. (3.17).

A.2.4 Phase invariant orthogonality and the Kronecker product

In this section it is proven that if s is phase invariant orthogonal, then $t = s \otimes v$ is also phase invariant orthogonal for any arbitrary row vector v. Let m be the length of v

and $k = pm + q$, $p = 0, \ldots, n-1$, $q = 0, \ldots, m-1$ then the following can be calculated:

$$\begin{aligned}
&\sum_{j=1}^{nm} t_{i,j} t_{i',1+(j+k)\%(nm)} \\
=&\sum_{j=1}^{n}\sum_{l=1}^{m} t_{i,m(j-1)+l} t_{i',1+(m(j-1)+l+k)\%(nm)} \\
=&\sum_{l=1}^{m} v_l v_{1+(l+q)\%m} \sum_{j=1}^{n} s_{i,j} s_{i',1+(j+p)\%n} \\
=&0 \qquad \forall i \neq i'
\end{aligned} \tag{A.43}$$

Eq. (A.43) has been used in Eq. (3.17).

Appendix B

Publications

B.1 Journal Publications

Emma Le Francois, Johannes Herrnsdorf, Jonathan J. D. McKendry, Laurence Broadbent, Glynn Wright, Martin D. Dawson and Michael J. Strain, *"Synchronization-free top-down illumination photometric stereo imaging using light-emitting diodes and a mobile device"*, *Optics Express*, vol. 29, no. 2, pp. 1502–1515, 2021

Emma Le Francois, Alexander D. Griffiths, Jonathan J. D. McKendry, Haochang Chen, David Day-Uei Li, Robert K. Henderson, Johannes Herrnsdorf, Martin D. Dawson and Michael J. Strain, *"Combining time of flight and photometric stereo imaging for 3D reconstruction of discontinuous scenes"*, *Optics Letters*, vol. 46, no. 15, pp. 3612–3615, 2021

B.2 Conference Submissions

Emma Le Francois, Johannes Herrnsdorf, Laurence Broadbent, Martin D. Dawson and Michael J. Strain, *"Top-down Illumination Photometric Stereo Imaging Using Light-Emitting Diodes and a Mobile Device"*, *OSA*, 2019

Emma Le Francois, Johannes Herrnsdorf, Jonathan J. D. McKendry, Laurence Broadbent, Martin D. Dawson and Michael J. Strain, *"Combined Time of Flight and Photometric Stereo Imaging for Surface Reconstruction"*, *IPC IEEE*, 2020

Research Article | Vol. 29, No. 2 / 18 January 2021 / Optics Express 1502
Optics EXPRESS

Synchronization-free top-down illumination photometric stereo imaging using light-emitting diodes and a mobile device

EMMA LE FRANCOIS,[1,*] JOHANNES HERRNSDORF,[1] JONATHAN J. D. MCKENDRY,[1] LAURENCE BROADBENT,[2] GLYNN WRIGHT,[2] MARTIN D. DAWSON,[1] AND MICHAEL J. STRAIN[1]

[1]*Institute of Photonics, Department of Physics, University of Strathclyde, Glasgow G1 1RD, UK*
[2]*Aralia Systems, Bristol Robotics Laboratory, Bristol BS16 1QY, UK*
[*]*emma.le-francois@strath.ac.uk*

Abstract: Three dimensional reconstruction of objects using a top-down illumination photometric stereo imaging setup and a hand-held mobile phone device is demonstrated. By employing binary encoded modulation of white light-emitting diodes for scene illumination, this method is compatible with standard lighting infrastructure and can be operated without the need for temporal synchronization of the light sources and camera. The three dimensional reconstruction is robust to unmodulated background light. An error of 2.69 mm is reported for an object imaged at a distance of 42 cm and with the dimensions of 48 mm. We also demonstrate the three dimensional reconstruction of a moving object with an effective off-line reconstruction rate of 25 fps.

1. Introduction

Photometric stereo (PS) imaging [1] is one of the most common 3D imaging methods for indoor scenarios. It can achieve better resolution than structured illumination imaging [2–4] or state of the art laser scanners [5], offers fast image computation [6], and it can deal with objects in motion and untextured area [7,8]. Compared to stereovision [9,10], only one camera needs to be calibrated, which reduces the computational reconstruction speed, the footprint and the cost [10]. PS imaging relies on having one fixed camera perspective and different illumination directions to image an object in 3D. This technique determines the surface normal vectors and surface albedo at each pixel of the captured images assuming a perfectly diffuse (Lambertian) surface of the imaged object [1]. Surface normal components can then be integrated to recover the 3D shape. Most work on PS has been developed combining various methods in addition to PS imaging, such as multi-view PS imaging [11], non-calibrated PS imaging [12] or self-calibrating PS imaging [13]. All the techniques report a good reconstruction accuracy, within millimeter range [11,12,14] and include real-time reconstruction [15].

Even though current work on PS imaging tackles major challenges such as uncalibrated PS imaging [13] and non-Lambertian PS imaging [16–19], most PS methods fail to demonstrate an easily deployable imaging technique that could show imaging applications in already existing building infrastructures. If this were achieved, PS imaging can provide an attractive route to using 3D imaging in industrial settings for process control and robot navigation, in public spaces for security and surveillance applications, and for structural monitoring.

There are two major obstacles that inhibit the widespread use of PS imaging for these purposes, which are the compatibility of the PS specific illumination with indoors or outdoors lighting installations, and the cabling required to synchronize several luminaires with each other and with the camera, which may potentially be mobile. Usually the camera and the luminaires are placed in the same plane, and a particularly common configuration employs four luminaires surrounding the

#408658 https://doi.org/10.1364/OE.408658
Journal © 2021 Received 3 Sep 2020; revised 25 Nov 2020; accepted 28 Nov 2020; published 11 Jan 2021

Optics EXPRESS

camera in a top/bottom/left/right or X-shaped configuation [5,8,14]. While such a setup is known to provide high-fidelity imaging results, it is incompatible with an application scenario where the luminaires are installed at the ceiling to provide room lighting, and a wall-mounted or mobile camera views the scene from the side (see Fig. 1(a)). Current PS systems use cables between the camera and the luminaires to enable synchronization, which is an undesired complication when retrofitting to existing lighting fixtures. Use of a WiFi or optically encoded clock signal can be a solution to remove the cabling, though additional infrastructure would be needed to implement this and the transmitters, camera and clock signal must be synchronized. Achieving synchronization using a "self-clocking" Manchester-encoded modulation scheme makes the approach described here easy to use, does not require additional infrastructure and can work in environment where WiFi is not available. Finally, traditional PS imaging often has a strong visual flicker and illumination low duty cycle, which is detrimental for indoors or outdoors lighting.

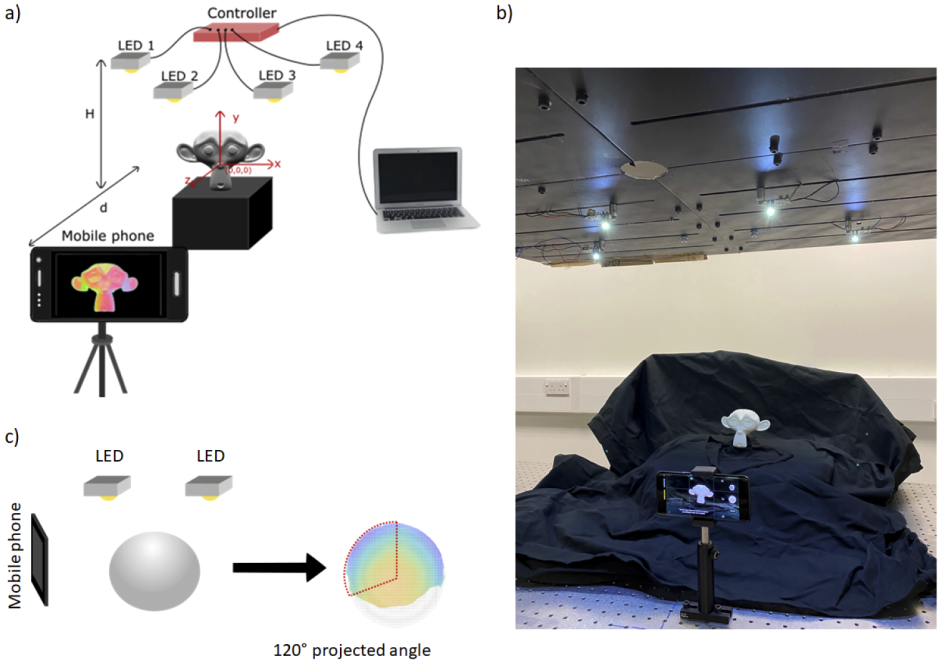


Fig. 1. Top-down illumination setup. a) Schematic of the photometric imaging setup, b) Picture of the photometric imaging setup and c) Schematic of the projected angle.

In this work, we present efforts to make PS imaging synchronization-free, reduce flicker, and demonstrate compatibility with ceiling lighting and both wall-mounted and mobile cameras. In this scenario, PS imaging would coexist with light fidelity (LiFi) networks [20] or visible light positioning (VLP) [21], potentially using the same light-emitting diode (LED) luminaires for all of these functions [22] as well as general lighting.

We demonstrate PS imaging using a hand-held mobile phone camera running at 960 fps and ceiling-mounted LEDs as illustrated in Fig. 1(a) and (b), operating in the presence of additional unmodulated lighting. Four LEDs operated by a controller board were mounted on a gantry were modulated with a bespoke binary multiple access (MA) format, referred to as Manchester-encoded binary frequency division multiple access (MEB-FDMA), that removes the need for synchronization and reduces flicker through Manchester encoding while maintaining a 50% duty cycle. A mobile phone set within the scene acquired a stack of frames at 960 fps, which were then processed to obtain the surface normal components, and finally integrated to obtain the

Optics EXPRESS

topography of the object. As the object is static, we do not reconstruct the full 3D object but rather its topography, commonly known as 2.5D reconstruction. For a static 48 mm diameter sphere, we report an root mean square error (RMSE see Eq. (8)) of 2.69 mm at a distance of 40 cm with an angle of reconstruction from the top-down illumination of 120°, which represents 78 % reconstruction of the surface of the sphere. Finally, we also demonstrate a dynamic imaging scheme, using a high-speed camera, where an ellipsoid rotates at 7.5 rotations per minute (RPM) and report an effective off-line 2.5D reconstruction of 25 fps.

2. Orthogonal LED modulation

One key feature in our setup is the removal of the requirement to synchronize LEDs with the camera or among each other, both of which is needed in the time division multiple access (TDMA) that is used in conventional PS imaging. This is achieved through MA, and frequency division multiple access (FDMA) has been used by the authors before to achieve unsynchronized PS imaging [23]. However, FDMA has some drawbacks, in particular strong perceived flicker since some of the LEDs have to be modulated at a fraction of the camera frame rate, and the sinusoidal modulation requires analog control of the LED brightness. Here, we use a bespoke modulation scheme, called Manchester-encoded binary FDMA (MEB-FDMA), which works with direct digital modulation of the LEDs, has significantly reduced flicker compared to FDMA, and keeps the advantage of not having to synchronize sources and camera. A comparison of MEB-FDMA with other MA schemes (FDMA, code division multiple access CDMA [24] , space division multiple access SDMA [25], wavelength division multiple access WDMA [26] and TDMA) that could be used for PS imaging is provided in Table 1. MEB-FDMA and some other MA schemes have the additional feature that they enable visible light positioning of receivers within the imaged scene through a relative signal strength approach (Sec. 3.3.4 in [21]).

Table 1. Comparison of MA schemes for *N* emitters

MA scheme	FDMA	CDMA	**MEB-FDMA**	SDMA	TDMA	WDMA
Modulation	real-valued	binary OOK	**binary OOK**	binary OOK	binary OOK	none
Duty cycle	50%	50%	**50%**	50%	$1/N$	100%
Frame length	$2N$	$2N$	$\mathbf{2^{N+1}}$	$2\log_2 N$ or $2\sqrt{N}$	N	1 (N fixed to 3)
Computation (flop/frame)	$\sim 10N\log_2 2N$	$4N^2 - N$	$\mathbf{(2^{3N+1} - 1)\sum_{i=1}^{N} 2^i}$	0	0	0
Synchronization required LED↔LED	No	Yes	**No**	No	Yes	No
Synchronization required LED↔receiver	No	Yes	**No**	Yes	Yes	No
VLP enabled	3D	3D	**3D**	2D	3D	3D
PS enabled	Yes	Yes	**Yes**	No	Yes	Yes
Visual perception	Poor (flicker)	Good	**Good**	Good	Poor (flicker)	Poor (monochrome luminaires)
Compatible with constant background light	Yes	Yes	**Yes**	Yes	No	No

Optics EXPRESS

2.1. Phase invariant orthogonal modulation

The important property that allows us to use FDMA without having to synchronize LEDs and camera is that if one frequency carrier experiences an arbitrary phase shift due to the lack of synchronization, it still remains orthogonal to the other frequency carriers. We call this property "phase-invariant orthogonality" and introduce it here formally before describing an alternative modulation scheme that shares this property.

Consider N emitters illuminating the scene over n discrete time steps. Then the $N \times n$ signal matrix $s_{i,j} \in \{-1, 1\}$ describes the time-sequence of on/off states of the individual LEDs. Here, $s_{i,j} = +/-1$ indicates that at time j the i^{th} LED element transmits a binary value of '1'/'0' in on-off keying (OOK), and after n binary values one 3D image frame is completed. Phase-invariant orthogonality requires that the rows of the matrix s remain orthogonal to each other even if they are time-shifted with respect to each other by an arbitrary phase Δj. Since camera pixels operate as integrating receivers, particularly at high camera frame rates, this requirement can be formalized to Eq. (1).

$$\begin{gathered}\sum_{j=1}^{n} s_{i,j}\left((1-\alpha)s_{i',1+(j-1+k)\%n} + \alpha s_{i',1+(j+k)\%n}\right) = 0 \\ \forall i \neq i',\ k = 1,\ldots,n,\ \alpha \in [0,1]\end{gathered} \tag{1}$$

Here the phase shift between rows i and i' is $\Delta j = k+\alpha$, and % is the modulo operator. Equation (1) represents the requirement from the experimental layout, however, mathematically it is equivalent to the simpler Eq. (2).

$$\begin{gathered}\sum_{j=1}^{n} s_{i,j}s_{i',1+(j-1+k)\%n} = 0 \\ \forall i \neq i',\ k = 1,\ldots,n\end{gathered} \tag{2}$$

FDMA with square wave carriers is phase invariant orthogonal. If s is phase-invariant orthogonal, then its Manchester-encoded version $s^{(1)}$ given by Eq. (3) - where $\otimes$ is the Kronecker product - is also phase-invariant orthogonal, *i.e.* all the benefits of Manchester encoding are readily available. Matrices of the form $s \otimes \begin{bmatrix}1 & 1 & \cdots & 1\end{bmatrix}$ are also phase-invariant orthogonal.

$$s^{(1)} = s \otimes \begin{bmatrix}-1 & 1\end{bmatrix} \tag{3}$$

Decoding of phase-invariant orthogonal encoded signals is less trivial than for CDMA schemes with synchronization. Equation (2) effectively means, that the operation of phase-shifting scatters the source signals into orthogonal sub-spaces of $\mathbb{R}^n$. Therefore, to enable successful decoding, the rows of the matrix $s_{i,j}$ need to be complemented by appropriately chosen orthonormal vectors $e^{(i)}_{k,j}$ that together with the rows of $s_{i,j}$ span all of these sub-spaces.

2.2. Manchester-encoded binary FDMA

We construct the MEB-FDMA carriers by starting with binary-valued square-wave FDMA. In order to be phase-invariant orthogonal over a sampling period T, the frequencies ν_i of the square waves must be in a fixed relationship given by Eq. (4).

$$\nu_i = \frac{p_i}{T} \qquad p_i \in \mathbb{N}^+ \tag{4}$$

A convenient choice of the integer values p_i is given by Eq. (5), in which $i = 1,\ldots,N$ identifies each LED:

$$p_i = 2^{i-1} \tag{5}$$

Optics EXPRESS

This means that the frame length $n^{(0)}$ without Manchester encoding is $n^{(0)} = 2^N$. We then construct a binary FDMA emitter signal $s^{(0)}_{i,j}$:

$$s^{(0)}_{i,j} = (-1)^{\lceil j/p_i \rceil}, \quad i = 1, \ldots, N \quad j = 1, \ldots, n^{(0)} \tag{6}$$

When using $s^{(0)}$, individual emitters may have long on and off times, leading to unacceptable visual flicker. Therefore, we use Manchester encoding:

$$\begin{aligned} s_{i,j} &= s^{(0)}_{i,j} \otimes \begin{bmatrix} -1 & 1 \end{bmatrix} \\ &= \begin{cases} (-1)^{\lceil j/2^i \rceil}, & j \text{ even} \\ (-1)^{1+\lceil (j+1)/2^i \rceil}, & j \text{ uneven} \end{cases} \end{aligned} \tag{7}$$

If the emitters are modulated with $s_{i,j}$ according to Eq. (7) using OOK, then they provide MEB-FDMA. A decoding algorithm for MEB-FDMA and underlying mathematical proofs are given in the appendix.

2.3. Properties of MEB-FDMA

For MEB-FDMA, similar to FDMA, any DC offset can be added to the received signal without affecting the decoding result. This is a prerequisite for applying the scheme to LED illumination since the intensity-modulated LED emission has only positive values. Furthermore, it allows installation of additional lighting fixtures that either do not carry a modulation signal or carry one at a much higher frequency, *e.g.* for LiFi.

Another remarkable property is that the transmitter and receiver can use the same sampling rate. This is surprising because the scheme uses Manchester encoding and the Nyquist theorem requires oversampling by a factor 2 to reliably identify each Manchester encoded bit. However, by requiring the modulation to fulfil the stringent criterion Eq. (1), the scheme was implicitly designed such that not every single Manchester bit needs to be identified individually. This property of the scheme means that the frequency of the LED modulation is the same as the camera frame rate and thus flicker is significantly reduced.

The number of OOK-bits needed for a single frame in MEB-FDMA scales exponentially with the number of emitters. Therefore, this modulation scheme is suitable for modest numbers of modulated emitters illuminating the camera field of view, typically 4-6 emitters in our suggested application.

3. 3D reconstruction process

The process flow of PS imaging in our setup is illustrated in Fig. 2, which comprises MEB-FDMA modulation and demodulation, PS processing and surface normal integration.

3.1. Frame acquisition and surface normal map

The transmitted signal is encoded as a clock signal, hence no trigger signal is needed to start the acquisition. Therefore, the acquisition simply starts when the recording button of the mobile phone is pressed. In practice, the LEDs have a 50% duty-cycle and thanks to the known optical fingerprint of each LED a decoding matrix can be created see Supplement 1. The received stack of images are therefore demodulated using the decoding matrix on each pixel of each image, see Fig. 2. At the end of the demodulation, four images corresponding to the four different illumination directions are retrieved.

The retrieved four images are then processed using established methods [1,14,27] to obtain the surface normal components N_x, N_y, N_z and the albedo A under the assumption of a Lambertian

Optics EXPRESS

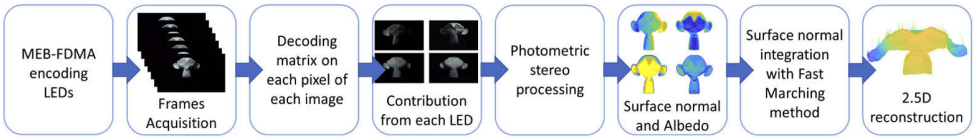


Fig. 2. Acquisition and Reconstruction program pipeline. LEDs are encoded with an MEB-FDMA scheme; the mobile device acquires a stack of images that are demodulated with a decoding matrix; four output images are then retrieved: one for each illumination direction; the photometric stereo processing determines the surface normal components and the albedo; afterwards the surface normals are integrated with a Fast Marching method to then obtain the 2.5D reconstruction of the object.

surface (see Fig. 2). As we are focusing our work on a new modulation scheme, we decided to use a conventional calibrated PS method to determine the surface normal map, hence the coordinates of each LED relative to the position of the object are needed to determine the lighting vectors.

3.2. Fast marching method

The next step of the reconstruction program is to integrate the surface normal vectors to obtain the topography of the object. Surface integration is a well known challenge and there are multiple methods in the literature for addressing it [28–30]. In this work, the surface normal vectors are integrated with the Fast Marching method [31–35] to take advantage of its reconstruction speed. The algorithm was implemented in Matlab and assessed on a data set from Yvain Queau [36], see details in the Supplement 1. The reconstruction process takes a few minutes to run on a desktop PC.

4. Optical acquisition system

As illustrated in Fig. 1, our system consists of a mobile phone device (Samsung Galaxy 9), four white LEDs (Osram OSTAR Stage LE RTUDW S2W) placed on a gantry above the object at a height of H = 46 cm, a controller board (Arduino Uno) for the LED modulation and a computer to communicate with the controller board and run the reconstruction program [37]. A series of geometric solids are 3D printed, namely a sphere with a 48 mm diameter, a cube which is 75 mm wide, and a complex shape of a monkey head that is 130 x 94.5 mm^2 wide and 79 mm deep. 3D printing ensures that the ground truth shape of the objects is known. On the setup, the geometric center of the object is the reference (0,0,0) and the location of the LEDs is determined from this reference point. The phone and the LEDs are located in two different planes. The phone is in front of the object on the z axis at a distance of d = 42 cm from (0,0,0) with a field of view (FOV) of 43 degrees. The LEDs are located at (x,y,z): LED1 (−27, 42, 10), LED2 (−14, 42, 35), LED3 (14, 42, 35) and LED4 (27, 42, 10), all in cm. The relative position of the LEDs to the camera and object positions have an impact on the extraction of the surface normal vectors. These coordinates are the best fit regarding the FOV of the scene and the object's size. A trade-off was made between the resolution of the reconstruction and the FOV. By placing the mobile phone at 42 cm from the object we can keep a mm-range depth resolution while assuming an orthogonal projection for the determination of the surface normal components. Moreover, a strict alignment of the mobile phone is not necessary in this work, as long as the FOV contains the front of the object then the orientation of the phone will not affect the accuracy of the surface normal components.

Each LED was modulated with an individual MEB-FDMA carrier signal at a on-off keying rate of 960 b/s. The phone captured frames with a resolution of 1280 x 720 at a rate of 960 fps for 0.2 s, with a black background to simplify image processing. The capture time is limited by the on-device data storage limit.

Optics EXPRESS

5. Results and discussion

5.1. Decoded frames

The decoded images of the three objects are displayed in Fig. 3. For the three cases, the light level clearly shows the different illumination directions. The brightness is slightly different depending on the position of the LEDs. This can be explained by the possibly imperfect match between the camera integration time and the MEB-FDMA scheme, i.e. the integration time may be shorter than the frame duration.

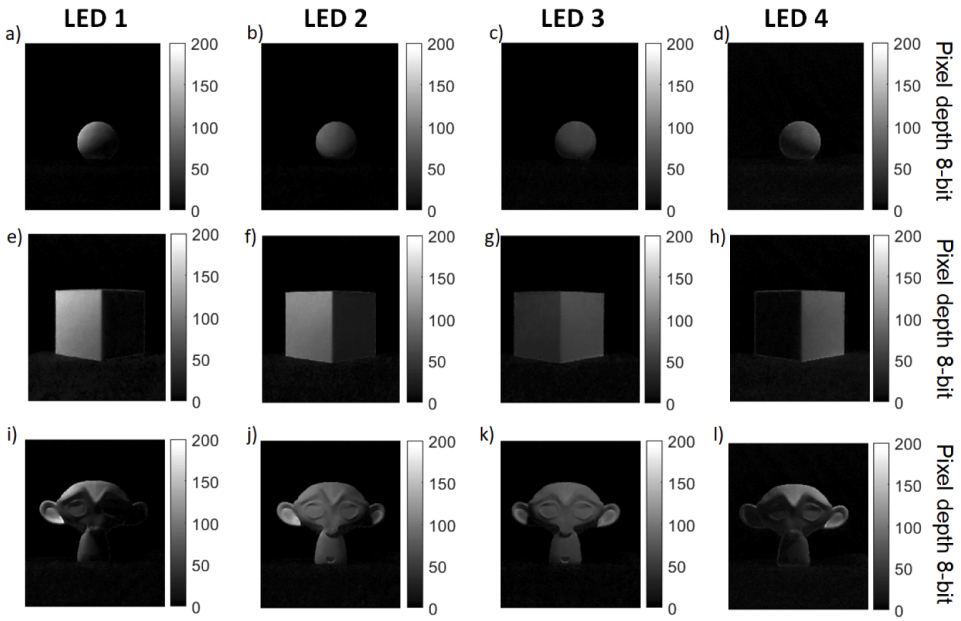


Fig. 3. Decoded images. Obtained after demodulation of the recorded frames for LED1, LED2, LED3 and LED4: a), b), c), d) for the sphere, e), f), g), h) for the cube corner and i), j), k), l) for the monkey head.

5.2. Surface normal vectors

From this set of images, surface normal components (N_x, N_y, N_z) and the albedo were calculated and are displayed in Fig. 4. For N_x, left and right facing surfaces are correctly distinguished as vector components with magnitude are ranging from -1 to 1. Similarly, N_y indicates up and down facing surfaces correctly, albeit with lower fidelity, and its value range is limited to -0.2 to 1 instead of -1 to 1. The poorer fidelity on N_y is due to the top-down illumination design as the bottom of each object is not suitably illuminated, which is also visible in the albedo plot. Moreover, as the camera is facing the object, N_z is positive and ranges from 0 to 1 with some variations due to the depth of the object. The albedo is normalized and is useful in understanding imperfections in the reconstruction. We notice that the albedo is more directional for the sphere and the cube corner than for the monkey, which is caused by the slight brightness variations seen in Fig. 3. Importantly, the surface normal components, which are the basis of the topography reconstruction, are observed to be less susceptible to these brightness variations than the albedo.

Optics EXPRESS

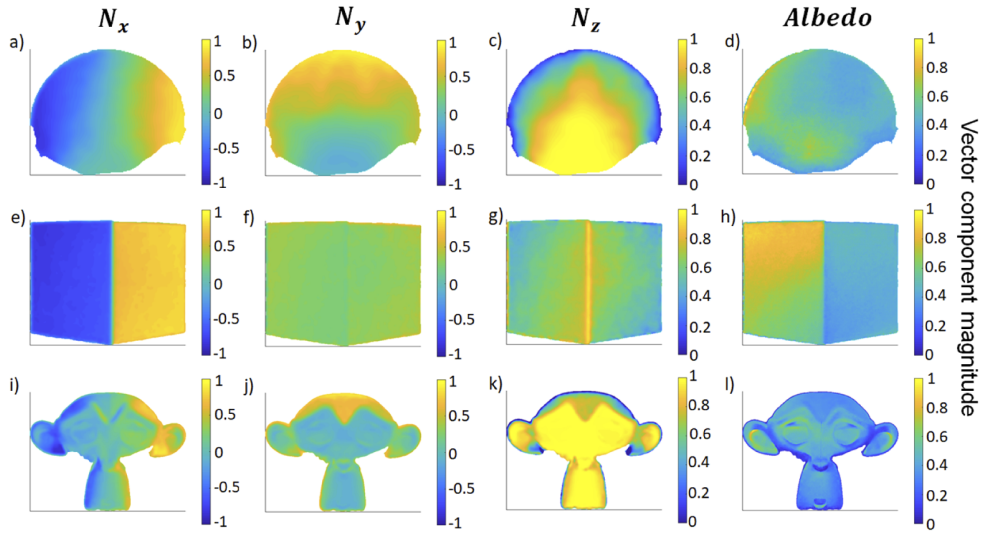


Fig. 4. Surface normal components and albedo. Obtained after running the photometric stereo algorithm, respectively N_x, N_y, N_z and Albedo: a), b), c), d) for the sphere, e), f), g), h) for the cuber corner, i), j), k), l) for the monkey head.

5.3. 3D reconstruction

Figure 5 and Fig. 6, respectively, plot the 2.5D reconstruction of the sphere, the cube corner and the monkey head in a perspective view, as well as a 3D rendered view (Blender). To render the reconstruction on Blender, a camera is set at a distance of 10 cm from the imported 2.5D reconstruction. For the sphere, Figs. 5(a)-(c) show a satisfactory global reconstruction for the top half of the object. The bottom half is poorly reconstructed and is "flat", which is also clearly shown on the rendered view. Because of the lack of information on the negative (downward facing) y axis, 78.4 % of the visible surface is reconstructed. The standard deviation is determined by the root mean square error (RMSE) and the normalised RMSE (NRMSE) which are defined as [14]:

$$RMSE = \sqrt{(\frac{1}{n}\sum_{i=1}^{n} z_i^2)}, \tag{8}$$

$$NRMSE = \frac{RMSE}{z_{max} - z_{min}}, \tag{9}$$

where n is the number of data pairs, z_i is the difference between measured depth values (along the z-axis) and reference values, and $(z_{max} - z_{min})$ is the range of measured values. According to the RMSE error map in Fig. 5(d), the most significant error is found at the bottom and on the edge of the sphere, while smaller errors in the top area are related to inaccuracies in N_z. Nonetheless, most of the error stays below 5 mm. Figure 1(c) shows the projected angle that can be retrieved using the top-down illumination setup where an angle of 120° is retrieved. An RMSE of 2.69 mm and an NRMSE of 5.61 % are obtained within the 78.4 % of the surface reconstructed. Despite the lack of information on downward facing facets, both standard deviation errors for the sphere are within the same range as in [14].

For the cube corner, despite the unequal partition of light on the cube and the important gradient variation, the reconstruction can retrieve the shape of the cube corner. Nonetheless, the top view shows that the reconstruction on the edge is deteriorating at the bottom of the object. This is explained with the top-down illumination configuration. Overall, by comparing

Optics EXPRESS

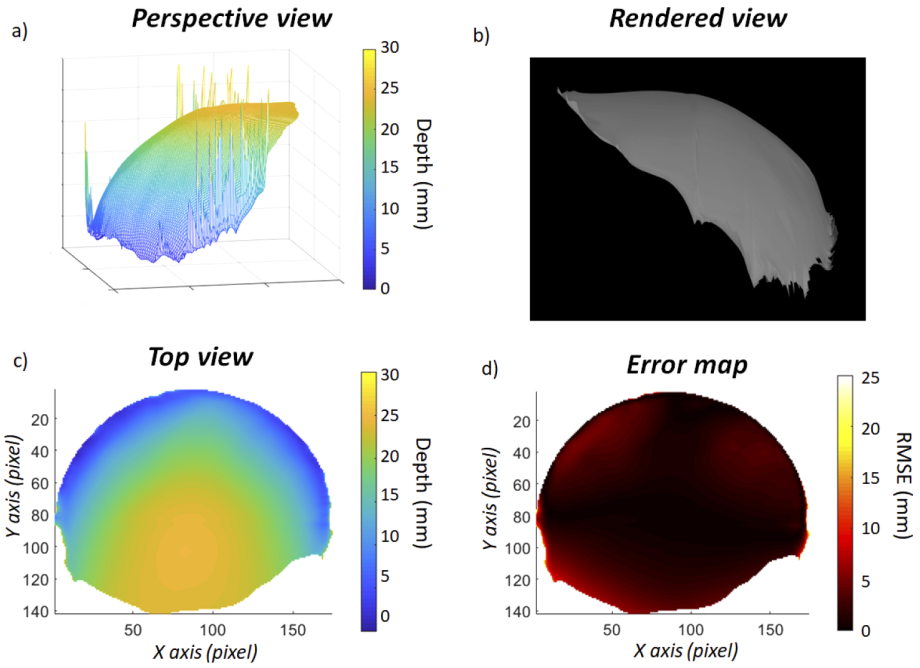


Fig. 5. 2.5D reconstruction of the sphere. a) Perspective view, b) Rendered view, c) Top view, d) RMSE error map.

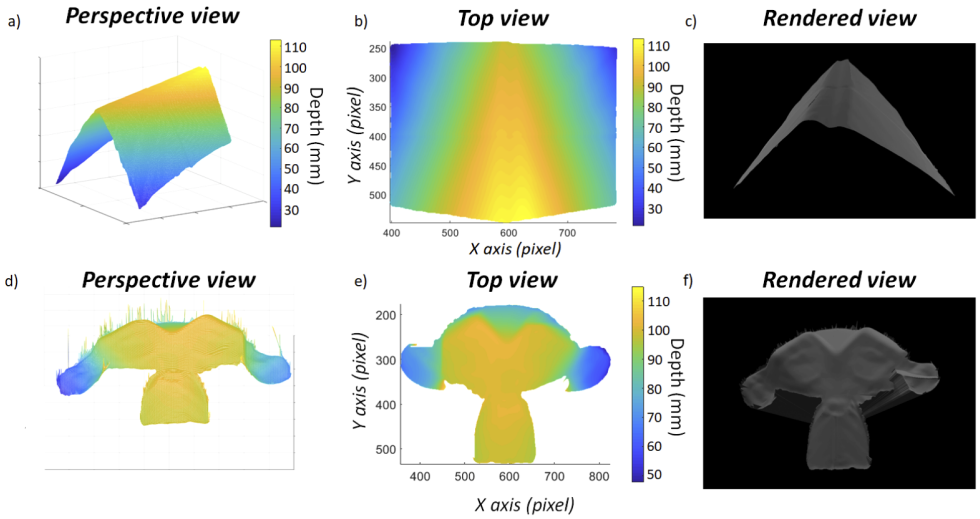


Fig. 6. 2.5D reconstruction of the cube corner and the monkey head. Cube corner: a) Perspective view, b) Top view, c) Rendered view. Monkey head: d) Perspective view, e) Top view, f) Rendered view.

Optics EXPRESS

the reconstruction of Figs. 6(a)-(c) with the images of Fig. 3 a good match is obtained. On the rendered view, an indent is observed and can be explained by the position of the starting point of the Fast Marching reconstruction process, from which the gradient values are integrated in a propagating fashion. At points with a strong gradient variation, the propagation will happen with very different gradient values in different directions and can create an indent on the row or the column of the starting point.

Finally, the monkey head has been chosen for its complex features, such as the eyes, the nose and the top of the head. The discontinuity between the face and the ears is a challenge for the Fast Marching algorithm to deal with. Indeed, the 2.5D reconstruction in Figs. 6(d)-(f) shows the shape of the nose, the eyes and also the upper head. However, the shape of the ears is harder to determine and the depth is substantially decreasing, which demonstrates the difficulty of the algorithm for dealing with those discontinuities. To quantify the error, a few features on the monkey face are measured: the horizontal size of an eye is 22 mm, the size of the nose is 1.1 mm and the distance between the eye orbits is 27 mm. After calibration, the same features are measured on the 2.5D reconstruction and we obtained the following measurements: 22.5 mm, 9.3 mm and 29.8 mm respectively. A few millimeters difference can be observed, which is close to the RMSE values obtained for the sphere. The top view reconstruction gives an idea of the percentage of the surface that is correctly reconstructed. The relative position between discontinuous regions is not handled well. However, features within each region are reproduced with good fidelity, such as the ears, on the rendered view. Small depth details, within mm, such as the earlobes and the eyes, are detectable and well reconstructed. Moreover, the distance between the face and the ears is about 50 mm and on the top view of the reconstruction the distance between the two features is also about 50 mm.

The 3D reconstruction relies on the surface normals, therefore most of the error on the reconstructed object topography will be dominated by the error on the surface normal vectors. Whenever N_z values are close to zero, the gradient integration during the Fast Marching process is numerically ill-conditioned. This phenomenon is clearly shown by the artefacts on the different reconstructions in Fig. 6.

5.4. Signal to noise measurement

Our modulation scheme can operate in the presence of additional unmodulated lighting. In order to assess its robustness, the reconstruction of the sphere is tested with different levels of background light in the room. For this experiment, the ceiling light of the room is illuminated and a voltage divider has been added on the LEDs in order to control the brightness and hence modify the signal power. After measuring the optical power of the signal and background light at the object, the SNR, in dB, is determined following Eq. (10):

$$SNR = 10 \log_{10} \left(\frac{P_{signal}}{P_{noise}} \right) \tag{10}$$

with P_{signal} the optical power in Watts of the LEDs and P_{noise} the optical power in Watts of the ceiling light.

Figure 7(a) shows a graph of the RMSE and the percentage of surface reconstructed versus the SNR for the sphere in the out-of-plane configuration. The SNR ranges from 0 *dB* to 5 *dB*. Across the SNR, the RMSE does not show a specific trend and the error ranges from 4 *mm* to 6 *mm* which is acceptable in our range of application. The percentage of surface reconstruction that is achieved over the measured range of SNR does show a dependance on the SNR. Figure 7(b) shows that as the SNR decreases, the reconstruction of the bottom part of the sphere is more and more challenging, which is a consequence of the top-down illumination configuration. This is explained by our reconstruction process. To avoid high error values being incorporated in the final image, a threshold is set 10 mm above the expected reconstruction value. Any values in the

Optics EXPRESS

final reconstruction above the threshold are discarded. This means that as the SNR decreases, we can reconstruct less area of the object, but the portion that is reconstructed is not affected by the SNR. Nonetheless, 65 % of the object view can be reconstructed even with an SNR just below 0 *dB*.

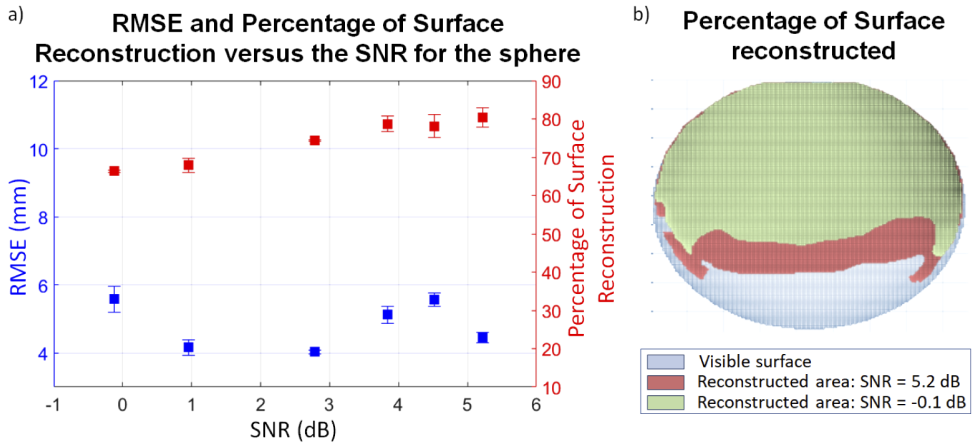


Fig. 7. Signal to noise result. a) Graph plotting the RMSE error and the percentage of the surface reconstructed regarding the signal to noise ratio for the sphere. b) Superposition of the different reconstruction of the sphere depending on the SNR.

5.5. Dynamic imaging

Finally, we reconstructed a moving object using our top-down illumination setup. A stepper motor (RS PRO Hybrid, Permanent Magnet Stepper Motor, 1.8° step) was added in order to rotate a 3D printed ellipsoid which was 100 mm long and 60 mm wide. A high-speed camera (Photron MiniUx100) replaced the mobile phone for its larger video capture memory thus enabling a longer acquisition time. The camera frame rate was set at 1000 fps with a shutter speed of 1 ms and was matched to the LED modulation rate. The stepper motor rotated at a speed of 7.5 RPM. Therefore, the acquisition ran for 8 s and the real-time video of the ellipsoid in motion can be found under the folder 'Dynamic imaging' in [38] (see Visualization 1). The 3D reconstruction is done off-line with the same reconstruction program pipeline as Fig. 2. The reconstruction requires at least 32 frames for a full reconstruction, and here we chose to record 40 camera frames for each 3D frame to match the motor step duration and to achieve effective full 3D reconstruction at a standard video rate of 25 fps.

Detailed analysis has been carried out on 21 representative 3D video frames, all separated by an angle of 18°. Therefore, our full 3D reconstruction relies on 21 topographies of the ellipsoid out of 200 possible reconstructions. To be able to clearly see the reconstruction and to match it with the display speed of the real-time video of the ellipsoid in motion, we decided to display the surface normal components and the reconstruction at 3 fps. This video display rate is 10 times slower than the effective rate we can achieve but it still represents a real-time 3D reconstruction video matching the object in motion. Two videos, one for the surface normal components and one for 3D reconstruction, can be found under the folder 'Dynamic imaging' in [38] (see Visualization 2 and Visualization 3 respectively).

The surface normals behave very similar to the static situation in that a high fidelity is observed in N_x, while N_y and N_z have lower but still useful fidelity. It is important to notice that the ellipsoid is constantly moving and despite its motion, its boundaries are sharp and well-defined which shows that the imaging rate is adequate.

Optics EXPRESS

In the 3D reconstruction video, the 21 reconstructed frames are repeated three times, showing both a color-coded plot of the reconstruction as well as a rendered view. Some errors can be observed at the edge of the ellipsoid which are artefacts caused by poor numerical condition due to $N_z \approx 0$. A flat reconstruction of the bottom of the object is also visible which is expected with the top-down illumination. Similarly to the sphere result with the static configuration, the ellipsoid is not entirely reconstructed. Some of the bottom part is missing which is not only due to the top-down illumination but also to the support piece that has been used to hold the ellipsoid in a tilted position. Nonetheless, the global 2.5D reconstruction of each view is satisfactory and of comparable quality to the static scenes.

6. Conclusion

In this work, we have been able to demonstrate an accurate 2.5D reconstruction of objects with different shape complexity with a RMSE error of 2.69 mm using a new photometric stereo imaging configuration that can readily be employed in conventional room lighting scenarios. We have also shown the 3D reconstruction of a moving object with an off-line effective 3D frame rate of 25 fps. Most importantly, MEB-FDMA encoding enables simple installation through removing the need for synchronization, and as importantly, modulates LEDs above the visual flicker recognition threshold, thus significantly simplifying the deployment, which was not possible before with successively flashed LEDs. Furthermore, we demonstrated that this method can be implemented using commercially available and hand-held mobile devices. Our work on synchronization-free top-down illumination photometric stereo imaging is currently at a proof-of-concept stage. However, future work will be focus on applying the method to digital lighting applications in public areas or industrial applications for surveillance, and also process and structural monitoring.

Funding

Engineering and Physical Sciences Research Council (EP/M01326X/1, EP/S001751/1); QuantIC (EP/T00097X/1); Fraunhofer UK (studentship of Emma Le Francois).

Acknowledgments

The development of the Fast Marching algorithm has been possible thank to Dr Juan Cardelino who shared his work on MathWorks File Exchange. We thank John Leck from Fraunhofer UK for his help on 3D printing the three different objects: sphere, cube and monkey head. We thank Dr Adam Polak for his help on the project. We thank Mark Stonehouse for his help on the experimental setup.

Disclosures

The authors declare no conflicts of interest.

See Supplement 1 for supporting content.

References

1. R. J. Woodham, "Photometric Method For Determining Surface Orientation From Multiple Images," Opt. Eng. **19**(1), 139–144 (1980).
2. D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 1 (2003).
3. M. Gupta, A. Agrawal, A. Veeraraghavan, and S. G. Narasimhan, "Structured light 3D scanning in the presence of global illumination," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* pp. 713–720 (2011).

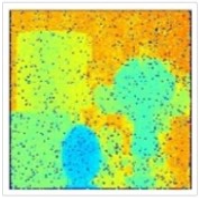
Optics EXPRESS

4. S. M. Haque, A. Chatterjee, and V. M. Govindu, "High quality photometric reconstruction using a depth camera," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* pp. 2283–2290 (2014).
5. Z. Lu, Y. W. Tai, F. Deng, M. Ben-Ezra, and M. S. Brown, "A 3D imaging framework based on high-resolution photometric-stereo and low-resolution depth," Int. J. Comput. Vis. **102**(1-3), 18–32 (2013).
6. J. Ackermann and M. Goesele, "A survey of photometric stereo techniques," Foundations Trends Comput. Graph. Vis. **9**(3-4), 149–254 (2015).
7. S. Herbort and C. Wöhler, "An introduction to image-based 3D surface reconstruction and a survey of photometric stereo methods," 3D Res. **2**(3), 4–17 (2011).
8. J. Herrnsdorf, L. Broadbent, G. C. Wright, M. D. Dawson, and M. J. Strain, "Video-Rate Photometric Stereo-Imaging with General Lighting Luminaires," *IEEE Photonics Conference* pp. 483–484 (2017).
9. A. Lipnickas and A. Knys, "A stereovision system for 3-D perception," Elektronika ir Elektrotechnika pp. 99–102 (2009).
10. S. E. Ghobadi, "Real time object recognition and tracking using 2D/3D images," Ph.D. thesis, University of Siegen (2010).
11. D. Vlasic, P. Peers, I. Baran, P. Debevec, J. Popović, S. Rusinkiewicz, and W. Matusik, "Dynamic shape capture using multi-view photometric stereo," ACM Trans. Graph. **28**(5), 1–11 (2009).
12. E. Salvador-Balaguer, P. Latorre-Carmona, C. Chabert, F. Pla, J. Lancis, and E. Tajahuerce, "Low-cost single-pixel 3D imaging by using an LED array," Opt. Express **26**(12), 15623 (2018).
13. G. Chen, K. Han, B. Shi, and Y. M. K.-y. K. Wong, "Self-calibrating Deep Photometric Stereo Networks," IEEE CVF CVPR pp. 8731–8739 (2019).
14. Y. Zhang, G. M. Gibson, R. Hay, R. W. Bowman, M. J. Padgett, and M. P. Edgar, "A fast 3D reconstruction system with a low-cost camera accessory," Sci. Rep. **5**(1), 10909 (2015).
15. G. Schindler, "Photometric Stereo via Computer Screen Lighting for Real-time Surface Reconstruction, International Symposium on 3D Data Processing," Visualization and Transmission pp. 1–6 (2008).
16. G. Chen, K. Han, B. Shi, Y. Matsushita, and K.-Y. K. Wong, "Deep Photometric Stereo for Non-Lambertian Surfaces," IEEE Transactions on Pattern Analysis and Machine Intelligence **8828**, (2020).
17. M. Li, C. yu Diao, D. qing Xu, W. Xing, and D. ming Lu, "A non-Lambertian photometric stereo under perspective projection," Front. Inf. Technol. Electron. Eng. **21**(8), 1191–1205 (2020).
18. K. H. Cheng and A. Kumar, "Revisiting outlier rejection approach for non-Lambertian photometric stereo," IEEE Trans. on Image Process. **28**(3), 1544–1555 (2019).
19. Q. Zheng, A. Kumar, B. Shi, and G. Pan, "Numerical reflectance compensation for non-Lambertian photometric stereo," IEEE Trans. on Image Process. **28**(7), 3177–3191 (2019).
20. H. Haas, L. Yin, Y. Wang, and C. Chen, "What is LiFi?" J. Lightwave Technol. **34**(6), 1533–1544 (2016).
21. T.-H. Do and M. Yoo, "An in-depth survey of visible light communication based positioning systems," Sensors **16**(5), 678 (2016).
22. J. Herrnsdorf, J. McKendry, M. Stonehouse, L. Broadbent, G. C. Wright, M. D. Dawson, and M. J. Strain, "Lighting as a service that provides simultaneous 3D imaging and optical wireless connectivity," in *2018 IEEE Photonics Conference (IPC)*, (2018), pp. 1–2.
23. J. Herrnsdorf, J. McKendry, M. Stonehouse, L. Broadbent, G. C. Wright, M. D. Dawson, and M. J. Strain, "LED-based photometric stereo-imaging employing frequency-division multiple access," in *2018 IEEE Photonics Conference (IPC)*, (2018), pp. 1–2.
24. J. K. Park, T. G. Woo, M. Kim, and J. T. Kim, "Hadamard Matrix Design for a Low-Cost Indoor Positioning System in Visible Light Communication," IEEE Photonics J. **9**(2), 1–10 (2017).
25. J. Herrnsdorf, M. J. Strain, E. Gu, R. K. Henderson, and M. D. Dawson, "Positioning and space-division multiple access enabled by structured illumination with light-emitting diodes," J. Lightwave Technol. **35**(12), 2339–2345 (2017).
26. C. Hernández, G. Vogiatzis, G. J. Brostow, B. Stenger, and R. Cipolla, "Non-rigid photometric stereo with colored light," *Proceedings of the IEEE International Conference on Computer Vision* pp. 1–8 (2007).
27. M. Chantler and J. Wu, "Rotation Invariant Classification of 3D Surface Texturesusing Photometric Stereo and Surface Magnitude Spectra," BMVC2000 pp. 49.1–49.10 (2013).
28. Y. Quéau, J. D. Durou, and J. F. Aujol, "Normal Integration: A Survey," J. Math. Imaging Vis. **60**(4), 576–593 (2018).
29. R. T. Frankot and R. Chellappa, "A Method for Enforcing Integrability in Shape from Shading Algorithms," IEEE Trans. Pattern Anal. Machine Intell. **10**(4), 439–451 (1988).
30. K. M. Lee and C.-C. Jay Kuo, "Surface reconstruction from photometric stereo images," J. Opt. Soc. Am. A **10**(5), 855 (1993).
31. J. Ho, J. Lim, M.-H. Yang, and D. Kiriegma, "Integrating surface normal vectors using fast marching method," Computer Vision-ECCV pp. 239–250 (2006).
32. J. A. Sethian, *Level Set Methods and Fast Marching Methods: evolving interfaces in computational geometry, fluid mechanics, computer vision, and material science*, Cambridge University (Cambridge monographs on applied and computational mathematics, 1999).

Optics EXPRESS

33. J. A. Sethian, "A fast marching level set method for monotonically advancing fronts," Proc. Natl. Acad. Sci. **93**(4), 1591–1595 (1996).
34. R. Sedgewick, *ALGORITHMS* (Addison-Wesley Publishing Company, 1983).
35. E. Ziegel, W. Press, B. Flannery, S. Teukolsky, and W. Vetterling, *Numerical Recipes: The Art of Scientific Computing* (Cambridge University, 1987).
36. Y. Quéau, J. D. Durou, and J. F. Aujol, "Variational Methods for Normal Integration," J. Math. Imaging Vis. **60**(4), 609–632 (2018).
37. E. L. Francois, J. Herrnsdorf, L. Broadbent, M. D. Dawson, and M. J. Strain, "Top-down illumination photometric stereo imaging using light-emitting diodes and a mobile device," in *Frontiers in Optics + Laser Science APS/DLS*, (Optical Society of America, 2019), p. JTu3A.106.
38. E. L. Francois, "https://doi.org/10.15129/44b337ab-9fd9-4983-9079-a0dcfaae1d84," DOI (2020).

Optics Letters Vol. 46, Issue 15, pp. 3612-3615 (2021) • https://doi.org/10.1364/OL.424000



Combining time of flight and photometric stereo imaging for 3D reconstruction of discontinuous scenes

Emma Le Francois, Alexander D. Griffiths, Jonathan J. D. McKendry, Haochang Chen, David Day-Uei Li, Robert K. Henderson, Johannes Herrnsdorf, Martin D. Dawson, and Michael J. Strain

Author Information ▾ Find other works by these authors ▾

Abstract

Time of flight and photometric stereo are two three-dimensional (3D) imaging techniques with complementary properties, where the former can achieve depth accuracy in discontinuous scenes, and the latter can reconstruct surfaces of objects with fine depth details and high spatial resolution. In this work, we demonstrate the surface reconstruction of complex 3D fields with discontinuity between objects by combining the two imaging methods. Using commercial LEDs, a single-photon avalanche diode camera, and a mobile phone device, high resolution of surface reconstruction is achieved with a RMS error of 6% for an object auto-selected from a scene imaged at a distance of 50 cm.

Full Article | PDF Article

More Like This

Synchronization-free top-down illumination photometric stereo imaging using light-emitting diodes and a mobile device
Emma Le Francois, Johannes Herrnsdorf, Jonathan J. D. McKendry, Laurence Broadbent,

Related Topics

Table of Contents Category
Imaging Systems

Optics & Photonics Topics
Image processing
Image reconstruction
Imaging techniques
Single photon avalanche diodes
Spatial resolution
Three dimensional imaging

About this Article

History
Original Manuscript: March 10, 2021
Revised Manuscript: May 21, 2021
Manuscript Accepted: June 22, 2021
Published: July 23, 2021

Top-down Illumination Photometric Stereo Imaging Using Light-Emitting Diodes and a Mobile Device

Emma Le Francois[1], Johannes Herrnsdorf[1], Laurence Broadbent[2], Martin D. Dawson[1] and Michael J. Strain[1]
[1]*Institute of Photonics, Department of Physics, University of Strathclyde, Glasgow G1 1RD (UK),*
[2]*Aralia Systems, Bristol Robotics Laboratory, Bristol BS16 1QY (UK)*
Email: emma.le-francois@strath.ac.uk

Abstract: 3D reconstruction of objects can be achieved using a top-down illumination, photometric stereo imaging configuration with four modulated white LEDs and a mobile phone. The Standard deviation for the reconstruction is ranging from 3.5% to 10.4%.

OCIS codes: 100.2000, 100.6890.

1. Introduction

Photometric stereo imaging relies on having one fixed camera perspective and different illumination directions to image an object in three dimension (3D) [1]. This technique determines the surface normal vectors and surface albedo at each pixel of the captured frames assuming a perfectly diffuse (Lambertian) surface of the imaged object [1]. Surface normal components can then be integrated to recover the 3D shape. So far, the most common photometric stereo configuration is as implemented with: a commercial camera placed in front of the object and, at least, four white light-emitting diodes (LEDs) surrounding it in a top/bottom/left/right or X shape [2,3]. A fast 3D reconstruction has been reported with white LEDs surrounding a camera, where LEDs were sequentially lit by a USB programmable board [2]. These 3D reconstructions showed a standard deviation error ranging from 2.65 mm to 15.60 mm for objects of size 50 mm and 160 mm, respectively [2].

In order to unlock the potential use of LEDs in industrial or public spaces for monitoring, security check or 3D imaging on mobile devices, the combination of lighting/camera system has to be easy to set up and flicker free. LEDs are affordable, energy efficient and have a fast modulation bandwidth ranging from several MHz up to GHz [4]. Thanks to a convenient interfacing with digital electronics [5], LED lighting enables advanced functionality such as wireless optical networking through the illumination itself.

Here we report the 3D reconstruction of objects using a top-down illumination photometric stereo imaging configuration. Four white modulated LEDs, mounted on a ceiling, illuminate the object while a mobile phone captures frames at 960 frame per second (fps). The LEDs are modulated at the camera frame rate with orthogonal multiple access carriers such that visible flicker is minimal and no electronic synchronization is needed between the LEDs and the phone. The 3D reconstruction shows a root mean square error (RMSE) ranging from 3.5% to 10.4% depending on the complexity of the object.

2. Experiment

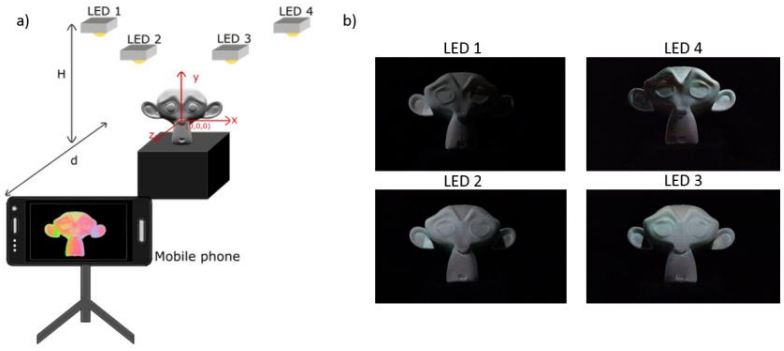


Fig. 1. a) Schematic of the experimental setup, b) Four images obtained from each fours LEDs after decoding

As illustrated in Fig. 1.a), four white LEDs (Osram OSTAR Stage LE RTUDW S2W) were placed on a gantry above the object at a height of H = 46 cm. In our experiments we used a number of 3D printed objects, including geometric solids and more complex shapes such as the monkey head in Fig. 1. The objects were ~130x94.5 mm wide and 79 mm deep, with the reference (0,0,0) point taken as their geometric centre. A mobile phone device

(Samsung Galaxy 9) was mounted on a tripod in front of the object on the z axis at a distance d = 30 cm with a field of view of 59 degrees. The phone was in a "side acquisition" configuration. The phone captured frames with a resolution of 1280x720 at a rate of 960 frame per second (fps) for 0.2 s. Fig. 1.b) shows the four images obtained after decoding the frames. For the illumination, we used an USB programmable controller board (Arduino Uno) to modulate the four LEDs at a frequency of 960 Hz. Each LED was modulated with an individual multiple access carrier signal at a frequency of 960 Hz, which is above visual flicker recognition and therefore suitable for digital lighting applications. The carriers were designed such that, analogous to orthogonal frequency division multiple access, no synchronization between the LEDs and the mobile phone was required.

3. Results

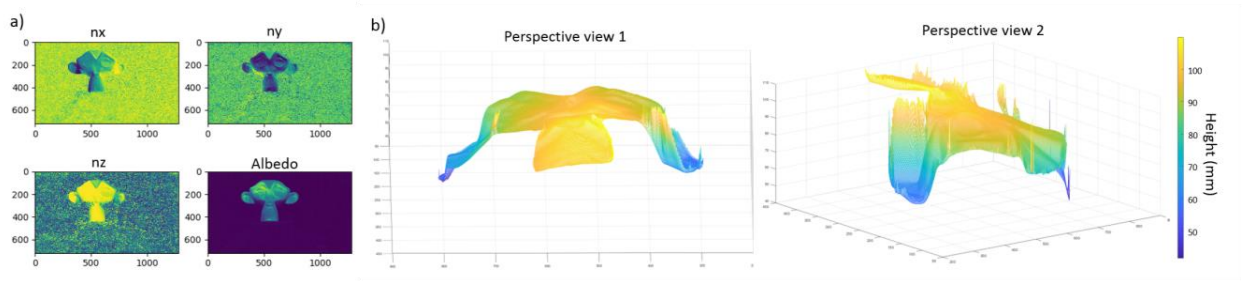


Fig. 2. a) Components of the surface normal vectors and the albedo, b) Different views of the monkey head 3D reconstruction

Surface normal integration has been a challenge for years in computer vision [6]. In this work, we integrate the surface normal vectors with the Fast Marching method [7] to take advantage of its reconstruction speed. A representative set of the surface normal components n_x, n_y, n_z, and the surface albedo are shown in Fig. 2.a). Yellow and blue color respectively represent positive and negative value of the surface normal components. As expected from the scheme in Fig. 1.a), n_x correctly distinguishes left and right facing surfaces of the object. Similarly, n_y correctly identifies up and down facing surfaces (clearly visible around the nose region), though we generally observe a poorer fidelity as compared to n_x due to the top-down illumination configuration. Finally, as we cannot see the back of the object, n_z is always positive with some variations due to the depth of the object.

Despite errors in n_x, n_y and n_z, especially in the ear area and the bottom of the face, 3D results plotted in Fig. 2.b) are comparable to the work done in [2]. As a reference, for a 48 mm diameter sphere and a 75 mm cube, in a similar configuration, we obtained a RMSE of respectively 5 mm and 2.6 mm which corresponds to the same RMSE range as [2]. The normalized RMSE of both objects equals to 10.4 % and 3.5 %, respectively.

4. Conclusion

Accurate 3D reconstruction is possible in a new photometric stereo configuration that can readily be employed in conventional room lighting scenarios. Top-down illumination photometric stereo can be directly applied to digital lighting application in public areas or industrial applications in the near future. Furthermore, we demonstrated that this method can then be implemented using commercially available, handheld mobile devices. Importantly, operation above the visual flicker recognition threshold is possible and there is no requirement for synchronization between LEDs, thus significantly simplifying installation and deployment.

5. References

[1] R. J. Woodham, "Photometric Method For Determining Surface Orientation From Multiple Images," Opt. Eng., vol. 19, no. 1, pp. 139–144, 1980.
[2] Y. Zhang, G. M. Gibson, R. Hay, R. W. Bowman, M. J. Padgett, and M. P. Edgar, "A fast 3D reconstruction system with a low-cost camera accessory," Sci. Rep., vol. 5, pp. 1–7, 2015.
[3] J. Herrnsdorf, L. Broadbent, G. C. Wright, M. D. Dawson, and M. J. Strain, "Video-Rate Photometric Stereo-Imaging with General Lighting Luminaires," pp. 483–484, 2017.
[4] J. Grubor, S. Randel, K. Langer, and J. W. Walewski, "Broadband Information Broadcasting Using LED-Based Interior Lighting," vol. 26, no. 24, pp. 3883–3892, 2008.
[5] J. J. D. McKendry et al., "Individually addressable AlInGaN micro-LED arrays with CMOS control and subnanosecond output pulses," IEEE Photonics Technol. Lett., vol. 21, no. 12, pp. 811–813, 2009.
[6] Y. Quéau, J. D. Durou, and J. F. Aujol, "Normal Integration: A Survey," J. Math. Imaging Vis., vol. 60, no. 4, pp. 576–593, 2018.
[7] J. Ho, J. Lim, M.-H. Yang, and D. Kriegman, "Integrating Surface Normal Vectors Using Fast Marching Method," Comput. Vision–ECCV 2006, pp. 239–250, 2006.

Combined Time of Flight and Photometric Stereo Imaging for Surface Reconstruction

Emma Le Francois[1], Johannes Herrnsdorf[1], Jonathan J. D. McKendry[1], Laurence Broadbent[2], Martin D. Dawson[1] and Michael J. Strain[1]
[1]Institute of Photonics, Department of Physics, SUPA, University of Strathclyde, Glasgow, UK
[2]Aralia Systems, Bristol Robotics Laboratory, Bristol, UK
emma.le-francois@strath.ac.uk

***Abstract*—3D reconstruction of objects can be achieved using both time of flight and photometric stereo imaging using four modulated white LEDs, a SPAD camera and a mobile phone. The standard deviation for the reconstruction is 4.1 mm at a distance of 70 cm.**

Keywords— Time of Flight, Photometric Stereo, 3D surface reconstruction, LEDs, SPAD camera, Mobile device

I. INTRODUCTION

Time-of-flight (ToF) and photometric stereo (PS) are three-dimensional (3D) imaging techniques with distinct, and complementary, areas of application. ToF imaging deals well with long range and discontinuous imaging but is limited by the resolution of current single photon camera systems or acquisition time of scanning systems. PS imaging uses conventional imagers and therefore has high spatial resolution but does not deal well with discontinuous surfaces. ToF is typically used for long-range light detection and ranging systems. A short-pulsed laser illuminates a scene in order to time correlate the reflected light intensity with the outgoing pulse to obtain a range map of the scene [1], and is commonly used in automotive and robotics applications. Alternatively, PS imaging is a passive method that relies on having one fixed camera perspective and different illumination directions to image an object in 3D [2]. This technique is more common in indoors scenarios for video surveillance, surface mapping and robot navigation [3]. Our previous work on "top-down" illumination PS imaging demonstrated 3D reconstruction with an error ranging from 3.5% to 10.4% for an object imaged at a distance of 42 cm [4]. However, this method used a black background to easily mask objects to reconstruct in order to speed up the computational reconstruction time. By employing a dual imaging system incorporating both ToF and PS the complementary properties of both systems can be used to image complex 3D fields with high resolution and complex discontinuities between objects.

Here we report the 3D reconstruction of an exemplar object (a sphere) using ToF - as a tool to mask the object, and PS imaging for a high-resolution surface reconstruction. A time-correlated single photon-counting (TCSPC) single photon avalanche diode (SPAD) camera is used with blue commercial light-emitting diodes (LEDs) to obtain a range map of the scene. For PS imaging, four white modulated LEDs illuminate the object while a mobile phone capture frames at 960 frames per second (fps). The LEDs are modulated at the camera frame rate with an orthogonal multiple access carrier schemes such that visible flicker is minimal, and no electronic synchronization is needed between the LEDs and the phone [4,5]. Our early results of the surface reconstruction show a root mean square error (RMSE) of 8.5%.

II. EXPERIMENT

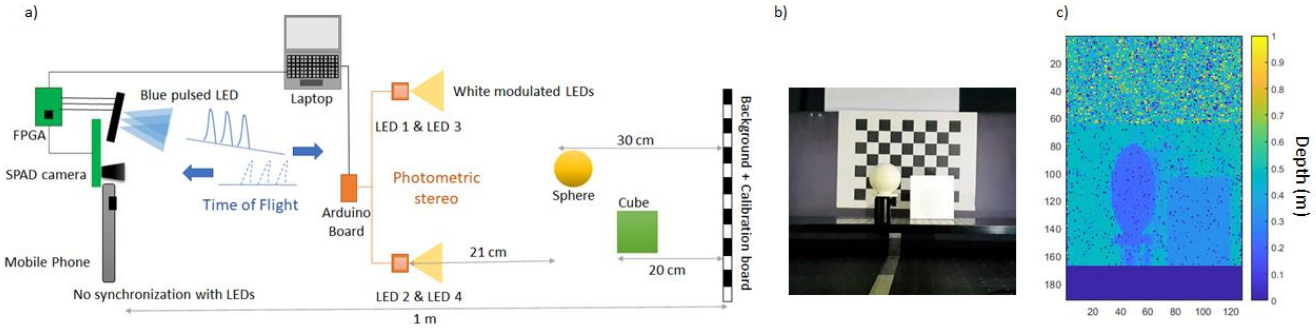


Fig. 1. a) Top-down schematic of the experimental setup, b) Cropped image taken with the mobile phone, c) Range Map of the scene using TCSPC mode with the SPAD camera

The experimental setup with both ToF and PS imaging is shown in Fig. 1.a). For ToF imaging, a SPAD image sensor is used and consists of a 192x128 SPAD pixels [5]. Each pixel is 18.4 x 9.2 μm² in area and can be operated with TCSPC functionality, see [5,6] for more details. Both photon counting (PC) and TCSPC modes are needed for the calibration. PC mode is used to acquire the image

978-1-7281-5891-4/20/$31.00 ©2020 IEEE

intensity of a calibration board and the TCSPC mode for ToF, where time signal is provided by blue commercial LED array that is pulsed – 11.3 ns pulse wide at a repetition rate of 30 ns – with respect to a trigger signal from the SPAD camera. Both LED array and SPAD camera are controlled with field-programmable gate array (FPGA) modules. For PS imaging, four white commercial LEDs were placed 21 cm away from the object in a X shape. A mobile phone device (Samsung Galaxy 9) was mounted on a tripod as close as possible to the SPAD camera at a distance of 70 cm from the object with a field of view of 32 degrees. The phone captured frames with a resolution of 1280 x 720 at a rate of 960 frames per second (fps) for 0.2 s (see picture in Fig. 1.b)). For the illumination, we used an USB programmable controller board (Arduino Uno) to modulate the four LEDs at a frequency of 960 Hz. Each LED was modulated with an individual multiple access carrier signal at a frequency of 960 Hz, which is above visual flicker recognition and therefore suitable for digital lighting applications. The carriers were designed such that, analogous to orthogonal frequency division multiple access, no synchronization between the LEDs and the mobile phone was required [4,5].

The SPAD intensity image was scaled and spatially registered to match the smartphone image dimensions using a checkerboard calibration object in the image. Then, a range map (Fig. 1.c)) obtained with the TCSPC mode is used to isolate the sphere by selecting its distance range. The subsequent PS imaging process carried out on the selected area follows the method from [4].

III. Results

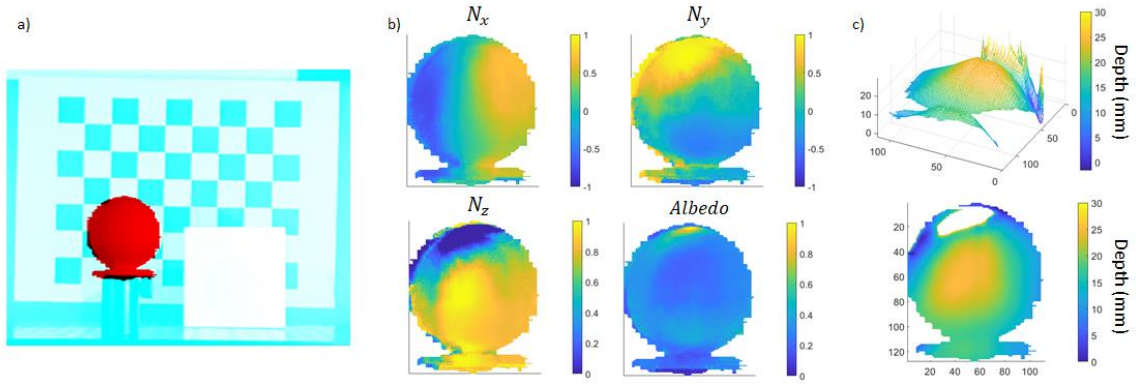


Fig. 2.a) ToF mask from SPAD camera superimpose on scene image, b) Surface normal components and Albedo of the sphere, c) 3D reconstruction of sphere.

Fig. 2.a) shows the ToF mask which is superimposed on the mobile phone image. We can see both cameras are correctly calibrated and aligned as the ToF mask coincides with the smartphone image of the sphere. Fig. 2.b) plots the surface normal components Nx, Ny and Nz of the sphere and its reflectivity (albedo). Nx and Ny correctly distinguishes left, right and up, down, respectively, facing surfaces of the object. As we cannot see the back of the sphere, Nz is always positive with some variations due to the depth of the object. The error apparent in Nz is likely due to the positionning of the four LEDs as the reflectivity is higher in this area, as shown in the albedo image. These surface normal components are used to calculate the surface topology with a reconstruction presented in Fig. 2.c). The error apparent in the Nz component clearly affects the top of the sphere reconstruction. Nonetheless, the global result is satisfactory and produces an RMSE of 4.1 mm, which represents a normalized error of 8.5%. This error is within our previous results [4] though the mobile phone is located at almost twice the distance in this setup.

IV. Conclusion

Accurate calibration of a SPAD camera with a mobile phone has been demonstrated which gives the possibility to use ToF to improve our previous PS imaging technique. By using both ToF and PS imaging, we could simultaneously obtain depth information of a public area for example with a high-resolution 3D reconstruction of selected elements within the imaged scene.

Acknowledgment

We acknowledge support from the EPSRC (EP/M01326X/1, EP/T00097X/1). Emma Le Francois's studentship was part supported by Fraunhofer UK. We would also like to thank David Li for lending us the SPAD camera. DOI: https://doi.org/10.15129/0e4cc5fc-2f8e-4ff9-98bc-bead040ef5ac

References

[1] M. Edgar, S. Johnson, D. Phillips, and M. Padgett, "Real-time computational photon-counting LiDAR," Optical Engineering, vol. 57, no. 03, p. 1, 2017.

[2] R. J. Woodham, "Photometric Method For Determining Surface Orientation From Multiple Images," Optical Engineering, vol. 19, no. 1, pp. 139–144, 1980.

[3] A. Lipnickas and A. Kny, "A stereovision system for 3-d perception," Elektronika ir Elektrotechnika, no. 3, pp. 99–102, 2009.

[4] E. L. Francois, J. Herrnsdorf, L. Broadbent, M. D. Dawson, and M. J. Strain, "Top-down illumination photometric stereo imaging using lightemitting diodes and a mobile device," in Frontiers in Optics + Laser Science APS/DLS, p. JTu3A.106, Optical Society of America, 2019.

[5] R. K. Henderson *et al.*, "A 192 × 128 Time Correlated SPAD Image Sensor in 40-nm CMOS Technology," *IEEE J. Solid-State Circuits*, vol. PP, pp. 1–10, 2019, doi: 10.1109/JSSC.2019.2905163.

[6] A. D. Griffiths *et al.*, "Multispectral time-of-flight imaging using light-emitting diodes," vol. 27, no. 24, pp. 35485–35498, 2019.

Appendix C

News release

Optica, News releases: *"Researchers Acquire 3D Images with LED Room Lighting and a Smartphone"*, 11 January 2021 [143]