Active Illumination for the Real World

Supreeth Achar

CMU-RI-TR-17-65

July 2017

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:

Srinivasa G. Narasimhan, Robotics Institute (Chair)
Simon Lucey, Robotics Institute
William L. "Red" Whittaker, Robotics Institute
Wolfgang Heidrich, KAUST and University of British Columbia
Kiriakos N. Kutulakos, University of Toronto

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

Keywords: Computational Imaging, Active Illumination, Depth Sensing

Abstract

Active illumination systems use a controllable light source and a light sensor to measure properties of a scene. For such a system to work reliably across a wide range of environments it must be able to handle the effects of global light transport, bright ambient light, interference from other active illumination devices, defocus, and scene motion.

The goal of this thesis is to develop computational techniques and hardware arrangements to make active illumination devices based on commodity-grade components that work under real world conditions. We aim to combine the robustness of a scanning laser rangefinder with the speed, measurement density, compactness, and economy of a consumer depth camera.

Towards this end, we have made four contributions. The first is a computational technique for compensating for the effects of motion while separating the direct and global components of illumination. The second is a method that combines triangulation and depth from illumination defocus cues to increase the working range of a projector-camera system. The third is a new active illumination device that can efficiently image the epipolar component of light transport between a source and sensor. The device can measure depth using active stereo or structured light and is robust to many global light transport effects. Most importantly, it works outdoors in bright sunlight despite using a low power source. Finally, we extend the proposed epipolar-only imaging technique to time-of-flight sensing and build a low-power sensor that is robust to sunlight, global illumination, multi-device interference, and camera shake.

We believe that the algorithms and sensors proposed and developed in this thesis could find applications in a diverse set of fields including mobile robotics, medical imaging, gesture recognition, and agriculture.

Acknowledgments

I am extremely grateful to my advisor Srinivasa Narasimhan for all the guidance and support he provided during my doctoral studies. The lessons I've learnt working along with him have had a great impact on my thinking and the way I approach problems. I couldn't have asked for a better mentor.

I would like to thank Red Whittaker, Simon Lucey, Kyros Kutulakos and Wolfgang Heidrich for making the time to be part of my thesis committee and for all the advice and feedback they gave me.

Parts of this thesis are the result of a very fruitful collaboration with Kyros and Matthew O'Toole from the University of Toronto and build upon some of their earlier work. Working with Matt has been a real privilege - he has great ideas and also the ability to develop them in a systematic, rigorous way. Over the last few years Kyros has been like an unofficial co-advisor for me - our weekly discussions and his insights have done a lot to shape this thesis.

Special thanks to Red who has been a wonderful advocate for the work developed in this thesis. He has supported and encouraged efforts to investigate applications for our sensors in robotics, mining and agriculture.

This work was done in the Illumination and Imaging Lab, a lab whose chief defect is that it is a gray, windowless room in the basement of Newell-Simon Hall. Its main redeeming feature is wonderful group of people working inside who have been my labmates in graduate school. Robert, Minh, Joe, Chao, Shumian, Jian and Suren have been great, intellectually stimulating company and they've always been willing to jump in and lend a hand while setting up experiments, collecting data or tracking down parts. Robert is also a co-inventor of FloorTennis™ and ScienceTime™, some of the finest afternoon diversions that have ever been created.

A lot of the work in this thesis required skills and knowledge that I lacked and which no one in our lab had either. Luckily, the Robotics Institute is a place where people are incredibly generous with both their time and their expertise. During the early stages of the Episcan3D project, Christoph Mertz and John Kozac showed me how they had accessed the synchronization signals of the projector we were using which helped the project along immensely. Then, of course, there is Chuck Whittaker. Before Chuck, our prototype was a jumble of cameras and cables held together by prayers and duct tape. After Chuck, Episcan3D was a cool red box whose content never needed to be fiddled with or realigned again.

My life in grad school was fun thanks to the company of a large and amazing group of friends. Abhijeet, Ajit, Apurva, Aranya, Aravindh, Arvind, Ashwati, Avinash, Chitra, Divya, Erle, Garg, Harsha, Harini, Ji, Jiuguang, Keshav, Lavanya, Maddali, Max, Mike, Nishtha, Riteja, Sahana, Sammy, Satya, Srinivas (the one who is not my advisor), Sudarshan, Suyash, Tommy, Toups, Uday, Utsav, Varun, Vas and Vinod - a big thank you to all of you. Pittsburgh wouldn't have been the same without the Youth Hostel, the zero Watt bulbs, the climbing, Dino, Mafia, or the "Table for 17 please" dinners and you are the people who made it all happen. The quizzing wasn't bad either.

Last, but definitely not least, I'd like to thank my parents and my sister without whose encouragement and support none of this would have been possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"You need to make something that works outside the lab."

Light and the way it propagates through an environment plays a pivotal role in how humans, animals and machines sense and perceive the world around them. One broad way of classifying visual sensing systems is to characterize them as being either passive or active. Passive visual systems have a light sensor but rely solely on external light sources present in the environment (typically sunlight) to provide illumination. In contrast, active visual systems have both a light sensor and a controllable light source that sends illumination into the environment.

Visual sensing in the biological world is almost exclusively passive. Most known instances of bioluminescence in nature are mechanisms to either lure prey, provide camouflage or provide means of communication. The deep-sea Black Dragonfish [99] is one of the few animals that emits light to aid its own sensing.

In man-made systems, the split between passive and active visual sensing is much more even. Passive vision systems can measure depth by stereo triangulation [63]. Additionally shading information [44] and defocus blur [35, 68, 80] contains cues that can be used to recover shape. Higher-level visual tasks such as tracking, segmentation and object recognition can be performed using passively captured images as inputs.

With passive visual systems, the image data captured is typically not sufficient to fully constrain the problem and provide a direct solution for the quantity (or quantities) of interest and so passive vision techniques generally rely on priors and inference to find a solution to these ill-posed problems. For instance- in passive stereo imaging, the goal is to generate a dense, complete depth map. Stereo correspondences are highly ambiguous in untextured parts of an image and so some sort of inference procedure is required to estimate depths in untextured regions based on the depths recovered at parts of the images with sufficient texture [87]. Similarly, for intrinsic imaging and shape from shading, a single image captured under one lighting condition is not enough to recover shape and reflectance so various priors need to be invoked to estimate a

solution [6, 55].

In contrast, active visual systems are generally less reliant on inference procedures and priors. In some cases, they can be designed to measure the quantity of interest directly. In many others, it is possible to capture enough visual data under one or more active lighting conditions and then find a unique solution by inverting a model of light propagation from the source to the sensor via the environment. The most common way of incorporating priors into active visual systems is to use methods from compressed sensing. These compressed sensing techniques are useful when capturing enough visual data to solve the problem uniquely would take too long [85], when the problem is still ill-posed despite being able to capture data with different lighting conditions [39].

1.1 Uses of Active Illumination

One large class of applications for active illumination is to capture images that are in some way better than the images that could be captured using purely passive techniques. A trivial but common instance is the use of a flash to capture photographs in low-light environments. Flash/no-flash image pairs can be combined [81] to obtain aesthetically pleasing, high quality images in low-light environments. In microscopy, many specimens are highly scattering (turbid) at optical wavelengths and so images captured under regular illumination have poor contrast and the parts of the specimen that are of interest can often not be seen clearly. There are a wide range of techniques for imaging through turbid that depend on active illumination [19]. Two prominent examples are scanning confocal microscopy [67] and light sheet microscopy [106] which both provide a means to optically section samples- that is capture light from a thin range of depths while blocking scattered and out-of-focus light from other depths.

Another very common use of active illumination is to build sensors for measuring depth. Different properties of light are useful for making depth measurements at different scales:

- Light travels at a constant speed in a medium (roughly $3 \times 10^8 \text{ms}^{-1}$). By measuring the time light takes to travel from an active source, to an object and back one can infer distance to the object. These time of flight (ToF) based active illumination methods can further be divided into two categories - impulse ToF (sometimes called direct ToF) and continuous-wave ToF (also known as indirect ToF). Time of flight is most useful for depth measurements at large scales - a few meters on upwards to a few kilometers.
- Light travels in straight lines through a medium. By finding the direction to a scene point from two known viewpoints, distance can be computed by triangulation. This is the basis for passive stereo depth sensing. With active illumination, triangulation is the basis of depth measurement using structured light and active stereo. The range and achievable resolution for triangulation-based techniques depends on the baseline separating the two

viewpoints.

- The wavelength of visible light ranges from 400nm to 700nm. For very small scale scenes (a few micrometers or smaller), the ray optics assumptions that triangulation depends on break down. However, at these scales, the wave nature of light can be exploited for shape measurements using interference based techniques. Examples of such techniques include optical coherence tomography (OCT) [45], phase shifting interferometry (PSI) [89] and coherence scanning profilometry [58]. In this thesis, we limit our attention to scales where geometric ray optics is valid and the effects of diffraction and interference can be ignored.

Closely related to depth measurement is the use of active illumination to measure shape. Photometric stereo [43, 108] estimates surface normals from the changes in brightness that occur when a scene is illuminated by controllable sources placed in different directions. Photometric stereo is the active illumination equivalent of shape from shading. Early photometric stereo techniques assumed that surface reflectance properties (the bidirectional reflectance distribution function or BRDF) were known, but current methods can jointly measure both shape and BRDF [24, 105]. Active illumination from multiple lighting directions can also be used to find depth discontinuities [82].

The most general way of describing light transport from a controllable light source (or set of sources), through an environment to a sensor is in terms of the light transport matrix [92]. The light transport matrix encodes all the information about geometry and reflectance that could be recovered from a set of light sources and sensors. However, directly measuring the light transport matrix for a projector-camera system is prohibitively expensive because of both the large size of the matrix and the high dynamic range of its elements. One way around this problem is to make assumptions about the form of the light transport matrix [85]. Another is to instead try measure (or 'probe') various types of components of light transport, such as direct and global illumination [69, 75], specular and diffuse reflections [54] or epipolar and non-epipolar light [76].

There are many other types of measurements that can be made with active illumination based sensing. For instance, the Doppler effect can be exploited to measure vibrations [95] and velocity [11, 41], active illumination sensing can be used to reason about parts of the environment that are outside a sensor's field of view [32] or measure the chemical composition of the atmosphere [7].

In this thesis, we look at how to develop hardware arrangements and computational techniques to make active illumination based sensing work robustly in a wide variety of difficult, real world conditions. We limit our attention to macro-scale scenes and focus primarily on the tasks of shape recovery and isolating elements of light transport.

1.2 Devices

Broadly speaking, active illumination depth sensors can be grouped into three classes - lidars, continuous wave time-of-flight cameras and triangulation-based devices. Here, we briefly explain the working principles of each type of device.

1.2.1 Lidar

Lidar (LIght Detection And Ranging) systems are impulse time-of-flight ranging devices. They measure distance by emitting a short pulse of light and measuring the time taken for the pulse's reflection off a scene point to return to the device. Because light travels at $3 \times 10^8 \text{ms}^{-1}$, the emitted burst of light must be extremely short (a few nanoseconds or less), the photosensor used to detect the reflected light must have large bandwidth and the clock used to measure the light pulse's round-trip time must have a very high resolution.

In contemporary lidar systems, avalanche photodiodes (APDs) [16] are usually the photosensor of choice. APDs have very fast response times and high light sensitivity which is important since the reflected pulse from a distant object can be weak. APDs can be operated in two modes - linear mode where the photodiode operates at moderate gain outputting a current proportional to the photon arrival rate and Geiger-mode where the gain is set very high causing the first detected photon to essentially saturate the photodiode. Geiger-mode APDs are also known as single-photon avalanche diodes (SPADs).

In a linear-mode lidar, the output of the photosensor can be processed to detect and timestamp the return using an analog timing discriminator or by using a high speed analog-to-digital converter to digitize the waveform for further processing. The latter approach has the advantage that it allows for handling multiple returns from the same direction. In a Geiger-mode lidar, the photosensor output is triggered at the first detected photon. Once a photon is detected, the photodiode needs to be reset before it can detect again. This means that at most one reflected photon can be detected per emitted light pulse. Since the photodiode can also be triggered by ambient light and dark counts (thermal noise), to obtain a reliable range measurement, multiple pulses need to be emitted and the returns need to be combined statistically.

Another important classification of lidar systems is on the basis of how large an area they sense at a time. Scanning lidar range finders have a emit light along one beam at a time and have a single photosensor, so they make depth measurements at a single point at a time. To create a depth image, the scan head needs to be actuated mechanically. Some scanning lidars contain an array of emitter-photodiode pairs that are all actuated together [33], but these devices are essentially just a set of point lidars packaged together. Flash lidars [57] emit light over an extended area at a time and have an array of photosensors, so they can capture at entire depthmap

at once instead of scanning point-by-point. The electronic circuitry each photodiode in a lidar needs for peak detection and time measurement is fairly complex, this makes fabricating sensor arrays for flash lidars difficult. A more detailed discussion of lidar principles and technology can be found in [64].

1.2.2 Continuous Wave Time-of-Flight Depth Cameras

Continuous wave Time-of-Flight depth cameras [56] emit light that is modulated temporally (typically in the $10\,\mathrm{MHz}$ to $200\,\mathrm{MHz}$ range). Depth can then be esimated (upto a modulo wrap around ambiguity) from the phase difference between the emitted light signal and the reflected signal that returns to the sensor from the scene.

To measure the phase of the reflected light signal, CW ToF cameras use sensor arrays use lock-in pixels that can redirect photoelectrons into different charge accumulation sites on the basis of the modulation signal. Readings from these accumulation sites can be used to demodulate the incoming light signal and recover its phase relative to the emitted signal (see section 5.2 for details). CW ToF cameras use lower intensity light sources than lidars and need to expose the sensor over multiple modulation cycles to be able to make accurate phase measurements. To integrate enough light at the sensor to recover a reliably phase measurement, CW ToF sensors typically use exposures that last for thousands of modulation cycles (ie. much longer exposure times than impulse ToF systems), but are able to generate depthmaps at video frame rates.

Unlike impulse ToF systems, CW ToF depth camera sensors can be fabricated with processes similar to those used to make regular CCD and CMOS image sensors (although SPAD-based implementations are also possible [73]) and do not have the extremely demanding bandwidth and timing requirements of impulse ToF systems like lidar. As a result, CW ToF sensors are cheap and easy to manufacture. The second version of the Microsoft Kinect is a popular consumer device that contains a CW ToF depth sensor [79].

1.2.3 Triangulation-Based Depth Sensors

Consider a camera and a controllable light source separated by a baseline. By emitting light in one direction from the source and then detecting the direction along which light reflected from the object reaches the sensor, the distance to the object can be triangulated. To speed up the acquisition process, light can be emitted along many directions from the source simultaneously to form a structured light pattern [94]. Images of one or more such patterns can be decoded to establish correspondences between directions along which light was emitted and received. Structured light devices typically use diffractive optical elements (DOEs) or programmable projectors to generate the projected patterns. The use of DOE light sources is common in consumer-grade

structured light systems aimed at gesture recognition like the original version of the Microsoft Kinect™ and the Intel RealSense™ F200. Projectors based on digital mirror devices (DMDs) can be used generate dense depth maps at extremely high framerates [60]. Structured light scanners are used in controlled industrial settings for quality control and similar tasks.

An alternative to structured light is active stereo where an active light source assists a stereo camera pair by projecting a spatial pattern onto the scene [52]. While structured light works only within the range of distances where reflected light from the active source can be detected by the camera, active stereo degrades gracefully to passive stereo at long ranges. The Intel RealSense™ R200 is an example of a active stereo device [50].

Continuous-wave ToF and structured light/active stereo based active illumination devices use low cost, easy to fabricate sensor technologies and so they would be well placed for wide scale adoption in a large number of application domains if they could be made to work robustly in real world conditions. We now look at some of the difficulties that currently limit their applicability.

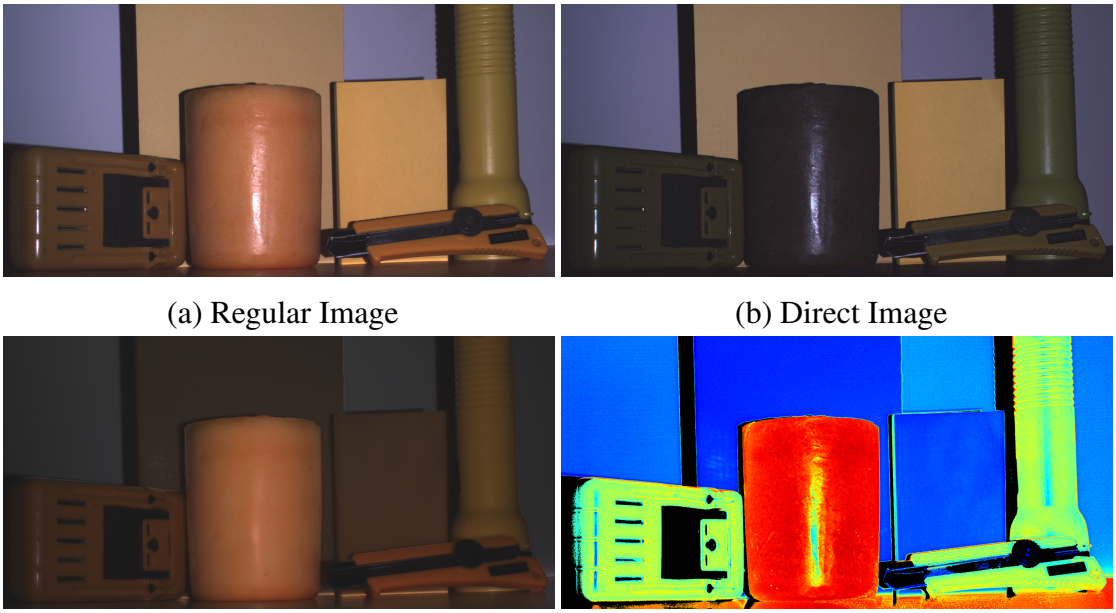
1.3 Challenges

Many materials and scene geometries interact with light in complex, difficult to model ways. Participating media such as smoke or fog scatter and attenuate light. The presence of bright ambient light sources (like the Sun) can make it difficult to reliably detect light from a low power active source. A system that uses lenses for illumination or imaging would need to be able to handle defocus in order to have a large working volume. To be useful in dynamic scenes or situations where the sensor is moving, an active illumination system would need to be able to correct for the effects of motion during the image acquisition process. We will now discuss each of these challenges in more detail and outline various approaches (both computational and in hardware design) that can be used to address them.

1.3.1 Global Light Transport

The radiance observed at a scene point illuminated by a light source is the sum of two components - the direct illumination and the global illumination. The direct component is the light that undergoes a single reflection in the scene along its path from the source to sensor. Light that takes indirect, multi-bounce paths forms the global component of illumination. The global component contains light that undergoes interreflections between surfaces, light that penetrates the surfaces of translucent objects (subsurface scattered light) and light that scatters volumetrically in a participating medium like fog or turbid water.

The direct component of illumination at a point is relatively easy to model. It depends on



(a) Regular Image (b) Direct Image

(c) Global Image (d) Fraction of Global Light

Figure 1.1: Direct and Global Illumination: A regular photograph of a set of objects made from different materials is shown in (a). Apart from some subtle variations in texture and specularities, there is little that can be used to differentiate between the different types of materials. Images (b) and (c) show the direct and global components of illumination in the scene separately. The waxy candle shows strong subsurface scattering so it appears very bright in the global image and only specularities and some surface details appear in the direct image. The yellow paper and cardboard show mostly surface reflection and so they appear dark in the global image. The plastic objects fall somewhere in between with some light appearing in the direct image and some in the global component. A simple statistic like the fraction contribution of global illumination to the total irradiance of a scene point (d) can encode useful material information.

the relative geometry between the light source and sensor, the position and orientation of the surface and the reflectance properties of the surface (which are described by the bidirectional reflectance distribution function or BRDF). The direct component at a point is independent of the configuration of the rest of the scene except for shadowing and occlusion effects. In contrast, the indirect component of illumination is usually difficult to model because the indirect component depends on the light a point receives from the rest of the scene.

Global illumination is pervasive in real-world scenes. Any scene with non-convex geometry or objects made of translucent materials will have global light transport - but because global illumination is so difficult to handle, active illumination techniques for shape recovery (both photometry-based and geometric) traditionally only account only the direct component of illu-

mination. With these direct-only models of light transport, global illumination in the scene acts as a noise source and can cause severe, systematic errors in depth estimates and recovered shape.

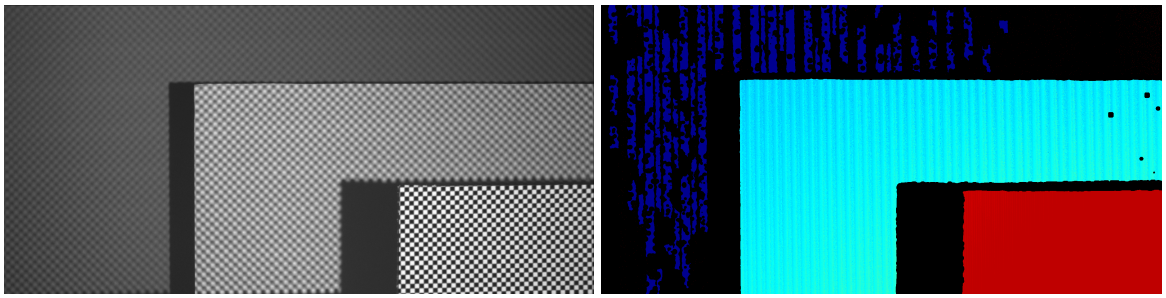
Prior work that looks at handling the global component in the context of active illumination can be broken into three groups. The first group of methods tries to explicitly model global light transport. An early example is Shape from Interreflections [71] which recovers the shape of concave Lambertian surface under distant lighting. The method iterates between shape estimation and correcting for the effects of interreflection on the estimated shape. Inverse light transport [91] generalizes this idea to arbitrary lighting conditions and develops interreflection cancellation operators that allow each n-bounce image in a scene with interreflections to be computed. With indirect ToF sensors, it is possible to combine measurements at different modulation frequencies with various simplifying assumptions about global light transport to correct for the effects of multi-path reflections [22, 31]

Another group of methods uses carefully designed active illumination patterns to avoid errors due to global illumination. It was observed in [69] that in most scenes, when the spatial frequency of the light pattern illuminating a scene is sufficiently high, the contribution of the global component to the radiance at a scene point becomes almost independent of the light pattern. In [14, 27] high frequency structured light patterns were designed to mitigate the effects of global illumination on structured light decoding. In [25] high frequency patterns were used for photometric stereo.

Thirdly, there are methods that block global illumination from reaching the sensor while imaging the direct component of illumination. This type of transport aware imaging was demonstrated in [75] where a camera was built that could probe just the global component of illumination. This probing technique was extended in [76] to capture images of arbitrary structured light patterns where global light transport's affects are constant. The key insight of [76] is that direct illumination always obeys the epipolar geometry between the light source and sensor while the majority of global light does not. In chapter 4 we describe a hardware arrangement that builds on this insight and can efficiently capture epipolar-only images of scenes illuminated by arbitrary structured light patterns.

While global light transport may appear to be a nuisance during shape recovery, being able to separately recover both the direct and global components of illumination can be useful. In addition to providing insights into how light interacts with a scene, having access to the two components separately instead of their sum can provide useful information about the physical properties of materials present in a scene. For instance, consider the scene in figure 1.1. All the objects are of similar colors, but when the components of illumination are separated using the technique proposed in [69] the difference between different types of materials becomes very clear.

1.3.2 Defocus



(a) Scene (b) Disparity Map

Figure 1.2: Illumination Defocus: In (a) a checkerboard pattern is projected onto a scene using a DMD projector. The scene consists of three planar targets at distances of 60cm, 80cm and 120cm from the projector-camera system. The pattern is focused on the nearest plane. Due to the shallow depth of field, the pattern is poorly focused at other distances. Running structured light on this scene yields poor results (b) because of illumination defocus. The foreground of the scene is reconstructed accurately because it is well focused, but the rest of the scene is reconstructed poorly.

Light sources and imaging sensors that use lenses to focus light have limited depths of field - a lens correctly focuses a single depth plane and only a narrow range of depths around this plane will have an acceptably small level of defocus blur. Thus, defocus can limit the working depth range of an active illumination system.

Illumination and imaging defocus blur can be reduced by using small apertures. This would increase depth of field but would also reduce the light throughput between the source and sensor. It is possible to use camera focus as a shape recovery cue (using depth from focus and depth from defocus methods [80]). It has been shown that in principle depth from defocus is similar to stereo triangulation based techniques but defocus cues have a baseline equal to the size of the aperture [88].

One of the issues with modeling camera defocus (both for recovering depth and all-in-focus images) is that the camera blur kernel at a pixel depends not only on the depth of the corresponding world point but also on the local surface geometry around the point. To reduce the complexity of the problem, most work in depth from camera focus/defocus makes a simplifying assumption - that the immediate neighborhood of any scene point is fronto-parallel to the camera. This assumption helps make the problem of inverting camera blur more tractable, but it comes at the cost of smoothing over fine details and introducing errors at depth discontinuities.

Projectors based on technologies like Digital Light Processing (DLP) and Liquid Crystal on

Silicon (LCoS) are frequently used as structured light sources for active illumination. These projectors typically use very large apertures to maximize light output. Illumination defocus differs from camera defocus in one crucial respect - the illumination defocus blur kernel at a projector pixel depends only on the scene depth of the corresponding world point and is independent of the geometry of the local neighborhood surrounding the point. As Zhang et al. observed in [110] this difference is because "projector defocus convolution happens on the projector's image plane while camera defocus convolution happens on the scene surface". In [110] illumination defocus was exploited for depth measurement and the technique was extended in [28] to handle both defocus and global light transport simultaneously.

Although defocus (in both illumination and imaging) can be used for shape recovery, defocus cues are typically less useful than stereo triangulation cues (the aperture of a lens is much smaller than the typical projector-camera baseline). Since illumination defocus blurs structured light patterns and camera blur mixes contributions from many scene points into a single image pixel, defocus is generally considered to be a error source in structured light. In chapter 3 we describe an illumination defocus-aware structured light technique that can extend the effective working volume of a projector-camera system.

Another solution to illumination defocus is to use a lensless light source to generate light patterns. Examples include diffractive optical element based systems and raster scanning laser projectors. In chapter 4 we describe a system that uses a scanning laser projector. In chapter 5 we develop a system that modulates light temporally instead of spatially and as a result is not affected by illumination defocus.

1.3.3 Ambient Light

Outdoors on a bright, sunny day the total irradiance due to the Sun observed on the surface of the Earth can reach as high as 1120Wm^{-2}. The controllable light sources used by active illumination systems are typically many orders of magnitude less powerful. For instance, a large tabletop projector would output roughly 10W of light, the active light source we use in chapter 4 has an effective output power of 20mW.

The arrival of photons at a sensor can be modeled as a Poisson process [104]. As the expected rate of photon arrival increases, the variability in the number of photons reaching the sensor over any time period also increases (if the expected number of photons is n, the standard deviation of the distribution is $\sqrt{n}$). When a scene point is illuminated by two sources, one weak and one strong, the signal due to the weaker source can get lost in the arrival noise (known as shot noise) of the stronger signal.

Shot noise fundamentally limits the ability of any light sensor to detect a weak light signal (like light due to a low power active source) in the presence of a far stronger light source (such

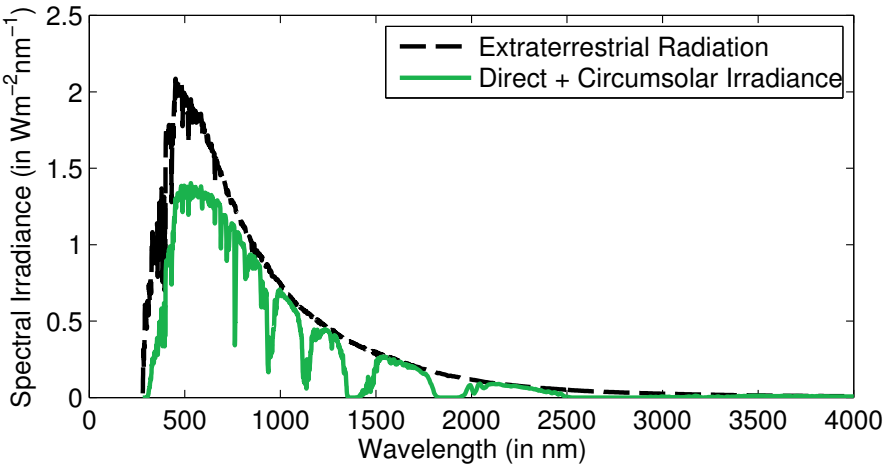


Figure 1.3: Spectral Distribution of Solar Irradiance: The black plot shows the spectral distribution of solar energy above the Earth's atmosphere. The green plot shows the spectral distribution at the Earth's surface. There are a few bands of wavelengths in the short wave infrared range which are almost completely absorbed by the atmosphere. Some high end laser range finders operate at these wavelengths but controllable light sources and sensors that work at these wavelengths are not well developed.(*Spectral data source: ASTM G-173)*

as the Sun). This problem is exacerbated by the r^2 fall-off in effective brightness of a point source with distance. Even when ambient light isn't strong enough to cause active illumination systems to fail completely, it adversely affects signal-to-noise ratio and degrades performance. In structured light techniques it can cause decoding errors and in modulated time-of-flight imaging it can increase noise in phase estimates.

Increasing the power of an active light source until it is bright enough to compete with sunlight is clearly not practical. The only way active illumination systems can work in bright ambient conditions is if they are designed so that a very large fraction of ambient illumination is blocked from reaching the sensor while all (or almost all) the light from the active source is allowed through.

The Sun is a wideband source that emits light over many wavelengths. A considerable fraction of sunlight can be blocked by using a narrowband active source and placing a matching optical filter in front of the sensor. The most effective optical filters are interference-based. They are designed for light incident normal to the filter and the passband shifts towards shorter wavelengths as the angle of incidence increases. This means that there is a trade-off between field-of-view and the selectivity of an interference-based optical filter that can be used for an imaging system. The wider the desired field-of-view, the broader the optical filter's passband needs to be.

As shown in figure 1.3 the spectral distribution of sunlight is peaked in the visible range, fairly strong in the near infrared (NIR) spectrum and tapers off at longer wavelengths. There are a number of wavelengths in the short wave infrared (SWIR) range where sunlight is weakened by the time it reaches the Earth's surface due to absorption in the atmosphere. It is possible to achieve some performance gains by operating at these wavelengths, but light sources and sensors that work in SWIR tend to be very expensive and have much poorer spatial and temporal resolution than their visible or NIR counterparts.

Almost all active illumination systems designed to be used outdoors use narrowband sources and wavelength filters but on its own, optical filtering generally does not block enough ambient illumination.

A complementary approach to handling ambient light is concentrating the active light source's output temporally and/or spatially. Scanning laser rangefinders and flash lidars emit extremely short, high power light pulses, but devices capable of generating, detecting and accurately timing light pulses at the short timescales needed tend to be very expensive.

There are some examples of spatial and temporal concentration being used in projector-camera systems. For instance, [65] demonstrated how a raster scanning projector can be used to build a low power line striping structured light system that works outdoors. In [30] a system is described that given a fixed power budget can vary the spatial concentration of light to reach a desired level of accuracy at the prevalent ambient light levels. In both these methods, acquisition time is sacrificed to gain robustness to ambient light. The Motion Contrast 3D scanner [13] breaks this trade-off by combining a raster scanning source with an event camera. In chapter 4 we describe a system that uses a raster scanning light source and which can image arbitrary spatial light patterns (not just single stripes) using an ordinary CMOS imaging sensor.

1.3.4 Interference

Since active illumination devices emit light into the environment, they can interfere with each other. Although the power of nearby interfering sources is likely to be far lower than ambient light levels, the effects can still be severe. This is because unlike ambient light, the light from interfering devices is modulated. Lidars are generally fairly robust to interference because the fraction of time over which the sensor is actually exposed awaiting the reflection of an emitted light pulse is very small, reducing the chance that two devices will interfere. Scanning lidars in particular are rarely affected by interference from other scanning lidars because they sense along a single direction at a time and the probability of two scanning lidars simultaneously pointing along the same ray or at the same point in the world is very low.

Interference from other devices is not a major concern in active stereo systems either because the interfering devices simply act as external sources of projected texture. Continuous wave ToF

cameras and structured light scanners however are strongly affected by interference from other devices. They typically expose their sensors at fairly high duty cycles and the decoding mechanisms they use can make gross errors (as opposed to just noisier measurements) if interference occurs. For these types of devices to be used widely in the real world it is important that multiple devices can operate simultaneously without severely degrading each other's performance.

1.3.5 Motion

Active illumination techniques can be divided into two classes based on the number of images they use. Multi-image techniques capture a sequence of images of a scene under different active lighting patterns while single-shot techniques use a single pattern and capture only one image. Multi-image techniques assume that an image pixel corresponds to the same world point across the sequence of images. This means that the light-source, camera and scene need to remain stationary during the acquisition process which is a severe limitation. For some tasks such as shape recovery using structured light or separating the direct and global components of illumination, both single-shot and multi-image methods exist while other tasks like flash/no flash photography [81] or capturing light-in-flight images with a ToF camera [38] inherently require multiple images to be captured.

Even when a single-shot solution to a problem exists, it would typically generate much lower quality results that a multi-image technique. For example multi-image structured light methods (such as gray codes [46] or phase shifting [94]) preserve fine scene details and do not make errors around depth discontinuities. In contrast, single-shot structured light methods ([48, 78]) produce lower quality depth estimates that are either sparse or have artifacts at depth discontinuities and in highly textured regions. This is because multi-shot structured light can decode correspondences temporally at each pixel while single-shot structured light requires a spatial window of support around each pixel to establish correspondence.

One way to address this trade-off between multi-image and single-shot techniques is to use spatio-temporal windows for matching correspondences as is done in spacetime stereo ([15, 111]). Since scene motion causes the shape of corresponding spatio-temporal windows to shear [111] searches for correspondences over both shifts (disparity) and shears (due to motion). Searching over shears effectively registers the images so that they appear to be images of a static scene taken from a fixed viewpoint. This generates high quality correspondences but is computationally expensive. Motion Aware Structured Light [101] uses structured light patterns that can be decoded both spatially and temporally and adapts the size and shape of the spatio-temporal window used for correspondence matching to handle scene motion.

Other approaches that attempt to compensate for the affects of motion in structured light are [51, 107]. Structured Light Transport [76] describes a method for interleaving multiple struc-

tured light patterns into a single image capture. Frames corresponding to each of the individual patterns can then be extracted from the captured image in a process similar to demosaicing. These extracted frames can then be passed through a regular multi-image structured light algorithm.

As discussed, when multiple images of a dynamic scene are captured, some sort of registration or motion compensation is required. During the capture of a single image, motion manifests as motion blur. If scene motion is small during the exposure time of the sensor, the effects of motion blur can be ignored. For indirect ToF sensors [59] describes a method for detecting pixel measurements that have been corrupted by motion blur. Direct ToF systems (both imaging and scanning lidars) use extremely short bursts of illumination and so they are unaffected by motion blur.

In chapter 2 we introduce a method for compensating for the effects of scene and sensor motion while performing direct-global separation using the multi-image high frequency illumination technique [69]. In chapter 5 we build a new hardware setup for epipolar-only imaging that can be used to capture images under different lighting conditions at very high speed effectively removing the need for motion compensation when using multi-image active illumination techniques on dynamic scenes.

1.4 Goals and Contributions

The goal of this thesis is to develop computational techniques and hardware arrangements to make active illumination devices that work under real world conditions. We want to combine the robustness of a laser scanning rangefinder with the speed, point density, compactness and economy of a consumer depth camera. Towards this end we make the following contributions in this thesis:

- **Motion Compensated Direct-Global Separation** [2] (Chapter 2): We have developed a method for registering images from a video stream where a dynamic scene is lit by a time varying projector pattern. This allows us to apply the direct-global separation method proposed in [69] to dynamic scenes captured from moving sensor platforms.
- **Multi-Focus Structured Light** [1] (Chapter 3): The working volume of active illumination systems that use lens-based projectors as a light source is severely constrained by projector's narrow depth of field. This limits the applicability of such systems to conditions where the relative positioning of the scene with respect to the system can be carefully controlled. We have developed a method that uses structured light patterns projected at multiple focus settings and combines illumination defocus cues with structured light triangulation to recover scene shape over large working volumes.
- **The Episcan3D Sensor** [77] (Chapter 4): We have developed a low power, portable active

	Motion Compensated Separation	Multi-Focus Structured Light	Episcan3D	Epipolar ToF
Global				
diffuse	✓	✓	✓	✓
subsurface	✓	✓	✓	✓
specular	✗	✗	✓	✓
volumetric	-	-	-	-
Ambient	✗	✗	✓	✓
Interference	✗	✗	-	✓
Motion				
dynamic scenes	✓	✗	✓	-
motion blur	✗	✗	✓	✓
Defocus				
illumination	✗	✓	✓	✓
imaging	✗	✗	-	-

Table 1.1: Comparison of techniques presented in this proposal on the basis of what types of effects they are robust to. Tick marks represent a high deal of robustness, hyphens are for effects that are partially handled and a cross indicates that the method is not robust to the corresponding effect.

illumination device that can capture the epipolar-only or indirect-only components of light transport at video frame rate with optimal light throughput. The epipolar-only imaging mode is robust not only to global light transport but also ambient light sources. The sensor can be used for structured light or active stereo depth sensing outdoors in bright sunlight. Our sensor illuminates and images one epipolar line at a time and strikes a balance between the need for light concentration and scanning speed.

- **The Epipolar Time-of-Flight Sensor** [3] (Chapter 5): The range of the Episcan3D sensor is limited by the projector-camera baseline and the extremely low power of its active light source. We extend the ideas behind the triangulation-based Episcan3D sensor to continuous wave time-of-flight imaging. Our Epitof (Epipolar Time-of-Flight) prototype inherits the robustness to ambient light and global light transport effects and has a range of $15\,\mathrm{m}$ outdoors in bright sunlight. Also, epipolar time-of-flight allows us to make depth measurements of fast moving scenes.

Chapter 2

Motion-Compensated Separation of Direct and Global Illumination

"When stereo gets out of line, they call it optical flow."

2.1 Introduction

Separating the direct and global components of illumination provides valuable insights into how light interacts with a scene. Being able to extract the direct component of illumination can improve the performance of classical photometry based algorithms like shape from shading as well as structured light reconstruction which typically do not account for global effects. Having access to the two components separately instead of their sum can also provide useful information about the physical properties of the materials present in a scene.

Efficient method for finding the global and direct components of illumination using source occluders or a projector-camera system were first proposed in [69]. The key insight of the chapter is that for most scene types, the global component of illumination varies fairly smoothly and has low spatial frequency. If a high spatial frequency pattern is projected onto a scene, the contribution of global illumination to the irradiance of each scene point will be independent of the light pattern being projected.

The method involves capturing images while illuminating the scene with a sequence of high frequency patterns. The separation can be performed using three sinusoid patterns, but the best results with practical projector-camera systems require around 20 high spatial frequency pattern images. A method that uses a single image was also presented, but it generates results at a fraction of the projector's resolution which is undesirable since most projector-camera systems are projector resolution limited. To generate high quality results, multiple images need to be captured under different illumination patterns and the light source, camera and scene need to

(a) Regular Image (b) Global Image

Figure 2.1: Motivating Example: Image (a) was captured using a regular flash, while image (b) is the global component of illumination computed using the separation technique of [69]. In the regular flash image it is difficult to tell which which bunches of grapes are ready for harvest. The global image (b), captures light that scatters below the surface of the grapes and the difference between bunches is much more pronounced. Being able to separate direct and global illumination from a moving platform would make it possible to visually scan a vineyard and determine what fraction of the crop was ready for harvest. (*Images courtesy Stephen T. Nuske)*

remain stationary during the image acquisition process.

In this chapter, we develop a method that relaxes the requirement that the scene and camera remain static during direct-global separation. This allows separation to be performed on video sequences in which the projector-camera system and/or the scene are moving and known time varying, high frequency patterns are being projected onto the scene. Figure 2.1 illustrates a motivating example for this problem. The method presented in this chapter makes no assumptions about the light source's ability to redistribute its output and can used with any type of projector that can project high spatial-frequency patterns onto the scene.

We assume that the underlying global and direct components of a scene point vary only slightly over small motions. This means that if the frames in a temporal window can be aligned, the separation technique in [69] can be applied to the aligned frames. Optical flow techniques can not be used directly because of the time varying patterns being projected onto the scene. Instead, we use a simple image formation model to approximate scene appearance under uniform lighting given an image of the scene under patterned illumination. We use these relit images to aid alignment and then estimate the global and direct components from the aligned images.

Compared to single image separation, our method produces more detailed, higher resolution results. We use all the frames in a temporal window for estimating the global and direct components. No frames are used exclusively for tracking, so our method can handle faster motions than

interleaving at a given frame rate. We use a colocated projector-camera system which allows us to avoid the difficult projector-camera pixel correspondence and 3D reconstruction problem.

We show that our method compensates for motion effectively and generates separation results close to ground truth. We show that not compensating for motion introduces significant artifacts in the separation and compare our method to alternatives such as single shot separation and interleaving. The scenes we demonstrate our method on contain materials such as wood, plastics, fabric, wax, leaves and human skin that interact with light in complex ways. We show that our method is able to compensate for both rigid motions and non-rigid deformations in the scene.

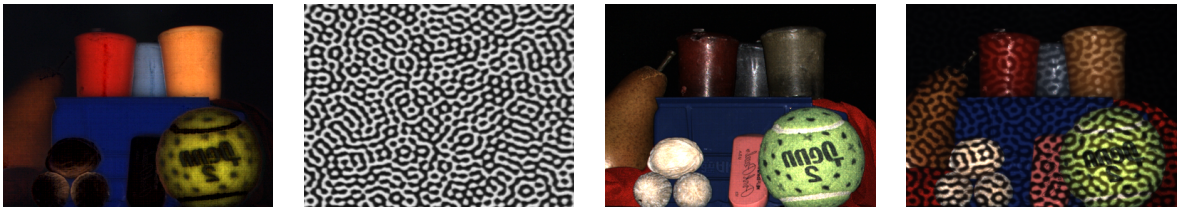
2.1.1 Related Work

By using a programmable mask in front of the camera and exploiting the high switching speed of DLP projectors, [75] was able to build a transport aware camera that could optically capture only the global component of illumination.

The need for motion compensation also arises in structured light for 3D estimation. In [107] a motion compensation method for the phase shift structured light algorithm is presented. Motion in the scene during image acquisition causes high frequency ripples in the phase estimates which are corrected by modeling the true phase as a locally linear function.

Motion estimation and compensation in image sequences with projected patterns is often done by interleaving the patterns with uniform lighting [112]. A similar approach is used in the structured light motion compensation scheme in [51] where patterns for structure estimation are interleaved with patterns optimized for estimating motion. Interleaving is a valid approach for our problem, but it increases the number of images that are needed and can introduce registration artifacts if the motion is not smooth.

Most techniques for optical flow are based on brightness and gradient constancy assumptions. Because we illuminate the scene with time varying, high frequency patterns, these assumptions are not valid. Illumination-robust optical flow methods have been designed based on photometric invariants [66] and physical models [37]. Computationally, most optical flow methods are based on local linearization of images. Since our images are dominated by the projected high frequency patterns, this linearization causes problems. An alternative optical flow formulation was derived in [96] that uses a direct search to compute optical flow and which can accommodate arbitrary data loss terms. We use a variant of this direct search method to refine our alignments.



(a) Global Component I_g (b) Projector Pattern s (c) Direct Component I_d (d) Observed Image I

Figure 2.2: Image Formation Model: The observed image I is modeled as a linear combination of the global component (I_g) and a projector pattern (s) modulated version of the direct component (I_d). $I = \frac{1}{2}I_g + sI_d$. The specularities on the candles appear in the direct image and most of the color is due to subsurface scattering in the wax and appears in the global image.

2.2 Image Formation Model

The brightness $I^t(x)$ of a pixel x at time t is a combination of the direct component I_d^t and global component of illumination I_g^t. When patterned illumination lights up the scene, the direct component is modulated by the pattern. If the pattern has an equal number of bright and dark pixels and has high spatial frequency compared to I_g^t, the contribution of the global illumination to the brightness is approximately $\frac{1}{2}I_g^t$ [69]. Thus we have

$$I^t(x) = \tfrac{1}{2}I_g^t(x) + s^t(x)I_d^t(x) \tag{2.1}$$

where, s^t is the value of the projected pattern at camera pixel x. We colocate our projector and camera so the mapping between projector and camera pixels is fixed and independent of scene geometry. Even though the projected patterns are binary, the value of s^t at a pixel can be continuous because real projectors do not have ideal step responses and the projector and camera pixels need not be aligned. An example of the model is illustrated in Figure 2.2.

For convenience and without loss of generality, we fix $t = 0$ to be the current frame, at the center of a temporal sliding window. As a scene point moves with respect to the projector and camera, its global and direct components will change. We assume that the motion within a sliding window is small enough for these changes to be negligible. This allows us to relate the global and direct components at time instant t in the sliding window to time 0

$$I_g^0(x) \approx I_g^t(W^t(x)) \quad I_d^0(x) \approx I_d^t(W^t(x))$$

where, W^t is an (unknown) warping function that aligns the view at time 0 to the view at time t. At $t = 0$, W is just an identity mapping. W^t depends on the geometry of the scene and the motion of the scene and projector-camera system. We use the notation $W \circ I$ to denote the image

that results from applying warp W to image I. It should be noted that the warp W^t is not an affine or projective mapping, but an a flow field over the image I^t with two degrees of freedom per pixel. Using this type of warp allows us to correctly handle rigid and non rigid motion in the scene.

Limitations The image formation model we use in this algorithm does not model changes in the underlying direct and global components at a scene point within a small temporal window. This is generally valid for global effects, diffuse reflections and smooth gloss reflections but not for sharp specularities. Additionally, we assume that there is no blurring of the images due to motion and that the scene lies completely within the depth-of-field of the projector-camera system. Additionally, the algorithm estimates W^t by computing optical flow and solves a number of total variation regularized optimization problems at each frame so it does not run in real time. The ability to register images togehter depends on the ability of the optical flow algorithm to generate correct flow fields. Thin structures, rapid motion and occlusions/disocclusions can adversely effect the performance of optical flow and hence affect our algorithm as well.

2.3 Motion Estimation and Compensation

We compute the direct-global separation at a frame in the video sequence using a small temporal sliding window centered at that frame. We seek to compensate for the motion that occurs inside a temporal sliding window so that the frames can be aligned to each other. With the help of the image formation model, we estimate how the scene would have appeared at each time instant under uniform lighting instead of the patterned illumination being projected onto the scene. We use these fully lit versions of the images as inputs to an optical flow technique that align the images. We then refine the warps generated by optical flow. Once the images are aligned we can compute the global and direct components robustly.

2.3.1 Estimating Scene Appearance Under Uniform Illumination

Finding the warps that align frames is difficult because of the time varying pattern being projected onto the scene. The time varying patterns violate the brightness and contrast constancy assumptions most optical flow methods rely on. To aid alignment, we compute an approximation of how the scene would have appeared ($\tilde{I}_f^t$) under uniform illumination from the frame I^t and the pattern s^t used to illuminate the scene. These fully lit estimates are better suited for image alignment than the original patterned frames.

Under uniform illumination, the brightness at a pixel is the sum of two unknowns, the direct component and the global component $I_f^t(x) = I_g^t(x) + I_d^t(x)$. The two unknowns are related by

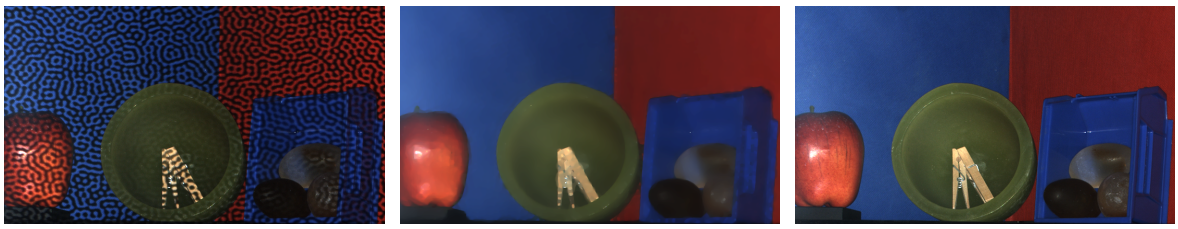
equation 2.1. The problem is under constrained and can not be solved uniquely because there is only one equation for every two unknowns. To find an approximate solution to the problem, we introduce a regularizer that enforces piecewise spatial continuity of the estimated global and direct components ($\tilde{I}_g^t$ and $\tilde{I}_d^t$ respectively). The loss function minimized is

$$L(\tilde{I}_g^t, \tilde{I}_d^t) = \|I^t - \tfrac{1}{2}\tilde{I}_g^t - s^t \tilde{I}_d^t\|_2^2 + \lambda_g TV(\tilde{I}_g^t) + \lambda_d TV(\tilde{I}_d^t) \tag{2.2}$$

where, λ_g and λ_d are smoothing parameters for the global and direct components. $TV(Z)$ is the isotropic total variation of the function $Z(x, y)$.

$$TV(Z) = \sum_{Domain(Z)} \sqrt{\left(\frac{\partial Z}{\partial x}\right)^2 + \left(\frac{\partial Z}{\partial y}\right)^2} \tag{2.3}$$

For color images, we sum together the total variations computed over each color channel. This objective function is similar to those used in $L1$ regularized image deblurring and denoising [10]. It is convex in the optimization variables $\tilde{I}_g$ and $\tilde{I}_d$ and can be solved efficiently to find a globally optimal solution. An example of the type of results found by our illumination pattern removal technique is shown in Figure 2.3.



(a) Input Frame $\frac{1}{2}I_g + sI_d$ (b) Recovered Fully Lit Frame $\tilde{I}_f = \tilde{I}_g + \tilde{I}_d$ (c) Actual Fully Lit Frame $I_f = I_g + I_d$

Figure 2.3: Recovering a Fully Lit Image: An image captured under a known high frequency illumination pattern (a) is relit to form an estimate (b) of how the scene would have appeared under uniform lighting. For reference, the actual appearance of the scene under uniform lighting is shown in (c). The relit estimate (b) captures the structure of the scene fairly well but smooths over some of the finer detail that is visible in (c) and the regularization introduces some artefacts (notice the pegs placed inside the green bowl). Relit images are used to perform image alignment on the original images

Any high frequency illumination pattern can be used to perform direct-global separation. We use random bandpass patterns similar to those used for structure estimation in [14]. The relit images sometimes contain minor artifacts along the edges in the patterns. These artifacts are caused by projector blur and small errors in the colocation between the projector and camera.

Using random patterns prevents correlation between these artifacts across time from affecting the registration process.

This procedure for estimating scene appearance under uniform lighting is in some ways similar to the image inpainting problem. In image inpainting, the goal is to 'repair' an image with missing regions by filling in the gaps with pixel values that agree well with the structure of their surroundings. Our goal is to 'repair' an image taken under patterned lighting to make it look like it was taken under uniform illumination. Approaches to the inpainting problem typically assume that images are self-similar [20] and/or exploit some type of sparsity structure [21]. The problem we solve here is similar to inpainting and we also make use of sparsity to help find a solution (the total variation regularizer enforces sparsity on the image gradients). There is one important difference though, in inpainting, pixels are typically marked as either good or bad and the goal is to fill in all the bad pixels. In contrast, in our problem each pixel has a continuous valued projector pattern associated with it instead of a binary label and we have an image formation model that relates projected pattern value to camera pixel brightness.

2.3.2 Registering Images

To align a frame to the center frame, we could simply compute optical flow between the relit frames. But for scene points that are illuminated in both images, it is better for the warp to match the original image pixel values $I^0(x)$ and $I^t(x)$ than the smoothed, relit estimates $\tilde{I}_f^0(x)$ and $\tilde{I}_f^t(x)$. At scene points that are not illuminated or which are illuminated in one frame but not the other, matching the relit estimates is preferable. We implement this idea in two stages. First, we compute the warping that best aligns $\tilde{I}_f^t$ to $\tilde{I}_f^0$ with variational optical flow [8]. This initial warp estimate is then refined by minimizing the following cost functional:

$$C(W^t) = \sum_x (1 - \alpha(x, W^t))|\tilde{I}_f^0(x) - \tilde{I}_f^t(W^t(x))| + \sum_x \alpha(x, W^t)|I^0(x) - I^t(W^t(x))| + \gamma TV(W^t) \quad (2.4)$$

where, $\alpha(x, W^t)$ is a weight that is high when a point is lit (s close to 1) in both the center frame I^0 and the current frame I^t. The total variation term is a regularizer to ensure that the computed warp is piecewise continuous. The weighing function α at each pixel is set as

$$\alpha(x, W^t) = \begin{cases} \left(s\left(x\right) s^t\left(W^t\left(x\right)\right)\right)^2, & \text{if } s\left(x\right) \geq 0.8 \text{ and } s^t\left(W^t\left(x\right)\right) \geq 0.8 \\ 0, & \text{otherwise} \end{cases} \quad (2.5)$$

The data term in the cost functional is a combination of the pixel differences between the relit images and the original images. The data term does not linearize well, so we minimize it approximately using the direct search algorithm proposed in [96]. Because we are using this step to correct small errors in an existing optical flow estimate we search for an refined warp at each pixel using a small window centered around the original warp estimate.

If the motion that occurs in a sliding window is large, optical flow may fail to correctly align some frames to the center frame. We detect poorly aligned frames by thresholding the correlation between the warped frame $W^t \circ \tilde{I}_f^t$ and center frame $\tilde{I}_f$. Poorly aligned frames are discarded from the sliding window. This allows the method to adapt to varying levels of motion in the scene. If motion is small and many frames can be aligned together reliably, they will all be used for computing the separation. As the speed and complexity of motion in the scene increases, the number of frames that can be successfully aligned to the center frame will reduce and the quality of the separation result will degrade gracefully.

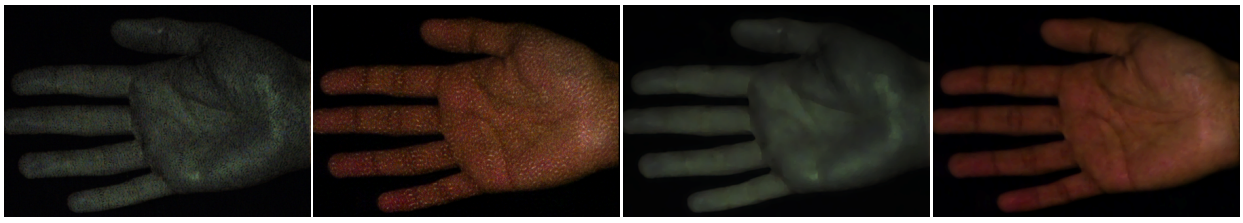
2.3.3 Computing Direct-Global Separation

Once the frames in a window have been warped to align with the center frame, we in effect have a set of images of the scene captured from the same viewpoint with different illumination patterns.

If the set of patterns is designed such that it can be guaranteed that for each camera pixel there will be at least one frame where the corresponding projector pixel is fully lit and another frame where it is fully dark, the separation can be performed using simple pixel-wise min and max operations over the frames in the sliding window [69]. Alternatively, since the projector pattern values (s^t) at each pixel are known, the global and direct components can be determined by fitting a line to the observed brightness values at a pixel using equation 2.1. For this line fit to make sense, each pixel needs to be observed under a range of projector pattern values.

Since scene structure and motion are not known a priori, a pattern sequence cannot be designed that is guaranteed to satisfy either of the above two criteria. As a result, there will be pixels in the image where the global and direct components can not be estimated well because the projector brightness did not change sufficiently at the corresponding scene point. We fill in these pixels by extending the idea of equation 2.2 to multiple images. We search for piecewise continuous global and direct components that are a good fit to the observed aligned image data by minimizing

$$L(I_g^0, I_d^0) = \sum_{t \in T} \| W^t \circ I^t - \tfrac{1}{2} I_g - \left(W^t \circ s^t \right) I_d^0 \|_2^2 + \lambda_g TV(I_g^0) + \lambda_d TV(I_d^0) \quad (2.6)$$



(a) Direct Image (b) Global Image (c) Direct Image (d) Global Image

Figure 2.4: Regularizing the Direct-Global Separation: Since scene motion is not not a priori, there will be pixels in the aligned frame stack that do not see a large range of projected light values. Separation using $\min$ and $\max$ operations or a linear fit would perform poorly at these pixels as can be seen in (a) and (b). By regularizing the total variation of the direct and global components, we can obtain much better separations as shown in (c) and (d).

where, T is the sliding window of frames selected about the center frame. At scene points where a variety of different s values were projected, $I_g^0(x)$ and $I_d^0(x)$ are estimated confidently as only a single line fits the data term. At pixels where the value of s was similar throughout the temporal window, many separations are plausible fits to the data and the smoothness terms help resolve the ambiguity.

2.4 Results

2.4.1 Experimental Setup

In our experiments, the scenes were illuminated using a 1024×768 DLP projector. For the experimental results presented in 2.4.2, the scenes were imaged with a Point Grey Grasshopper camera at 10 frames per second. For the results on deformable objects (2.4.3), the scenes were acquired at 60 frames per second. For all experiments, the camera was radiometrically calibrated to have a linear response curve and the camera and projector were colocated using a plate beam splitter. The camera used had a Bayer array for capturing RGB images (as opposed to being a 3-CCD sensor). This Bayer array reduces the effective resolution of the camera, but the effective resolution was still slightly higher than that of the projected patterns. Each raw camera image was debayered and the 3 color channels were processed independently. The experimental setup is shown in Figure 2.5.

An offline calibration step is used to find the image s corresponding to each illumination pattern. Each pattern is projected onto a diffuse, planar white surface and imaged by the camera. To correct for projector vignetting, all images were normalized with respect to a reference image

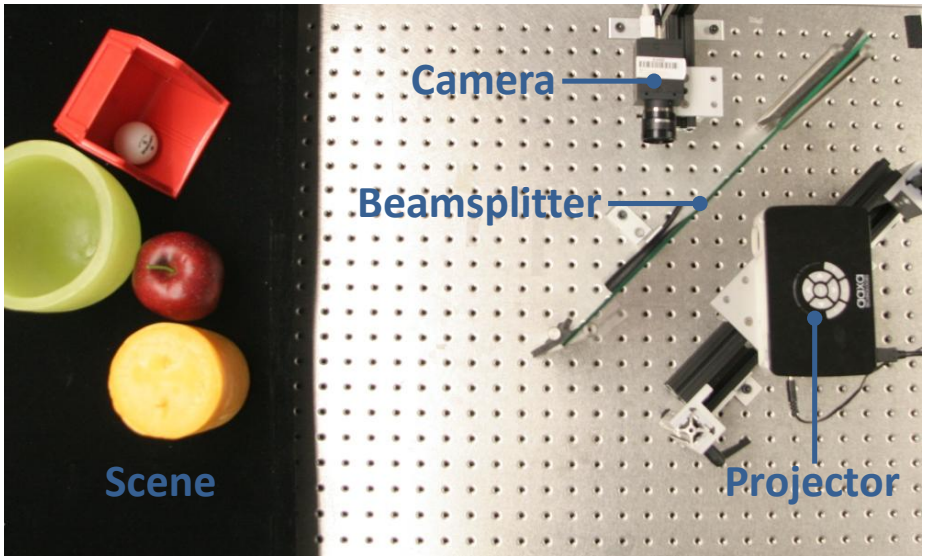


Figure 2.5: Experimental Setup: Our experiments were performed using a camera and projector colocated with a plate beam splitter. We also have a portable version of the setup.

of the same planar surface while fully lit by the projector. This calibration needs to be performed only once.

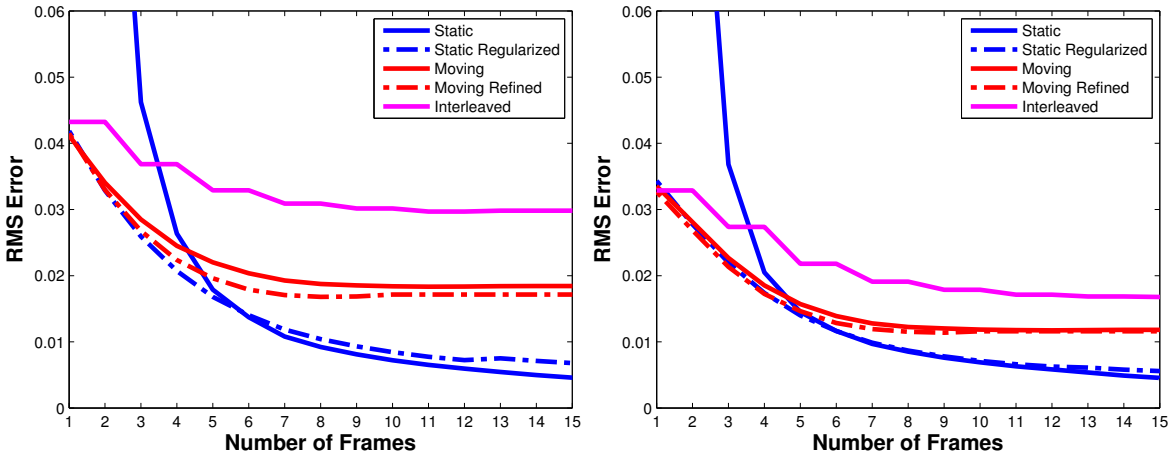
The regularization parameters λ_g and λ_d were tuned by hand. For the experimental results presented below, both λ_g and λ_d were set to 0.1 during the initial estimation step (equation 2.2). During the refinement step (equation 2.6), λ_g was set to 0.08 and λ_d was set to 0.04. The choice of the regularization parameters depends on the scale of the projected patterns and the scene albedo/texture as seen by the camera. Setting the regularization term values too high results in over smoothed results and setting the values too low results in noisy results where the projected pattern leaks into the separation results.

2.4.2 Comparisons on Rigidly Moving Scenes

The goal of these experiments is to compare the direct and global components generated by our algorithm on moving scenes to ground truth and to analyze the effect of temporal window size on separation accuracy.

Ground truth was acquired by first capturing 25 frames of a scene while projecting checkerboard patterns at different offsets. These frames were captured while the scene and camera were stationary. The direct and global components calculated on these 25 frames are used as ground truth (RMS Error 0). We then captured a video sequence with the scene in motion while patterns were being projected. The pose of the first frame in the video matches the pose from which the ground truth frames were acquired.

This experiment was performed from two different poses on two scenes (see Figs. 2.3 and



(a) RMS Error in Direct Component

(b) RMS Error in Global Component

Figure 2.6: Comparison of Separation Techniques: The RMS errors in the direct (a) and global (b) components for different separation techniques as the number of frames used is varied. The blue 'static' curves are from direct-global separation on stationary scenes and represent the best possible performance a method could achieve for a given number of frames. The red 'moving' curves are from using our motion compensation algorithm on moving scenes. When the number of frames is small, the motion compensation method performs just as well on the moving sequences as normal separation on an equal number of static frames. When the window size increases, frames far away from the window center are discarded because alignment fails and so performance of the motion compensated algorithm levels off.

2.7) yielding a total of four trials. RMS errors against ground truth for different separation methods averaged over the four trials are shown in Figure 2.6.

In the static case, the results improve in quality as the number of frames used increases ('Static' in Fig. 2.6). We also used the regularization method described in 2.3.3 on the static sequence ('Static regularized' in Fig. 2.6). The regularization improves performance when the number of frames is small and many pixels have not seen enough different projector pattern values. However, it smooths over some of the fine details in the scene and so it does not perform as well the unregularized technique when more frames are used.

For the video sequence corresponding to each trial, we tested our motion compensation method with different sliding window sizes using the first frame as the window center. We evaluated the motion compensation with the warp refinement described in 2.3.2 ('Moving Refined' in Fig. 2.6) and with the unrefined warps ('Moving' in Fig. 2.6). We also tested an interleaved approach where the projector alternates between patterns and uniform illumination ('Interleaved' in Fig. 2.6). Warps computed between the fully lit images are interpolated to align the patterned

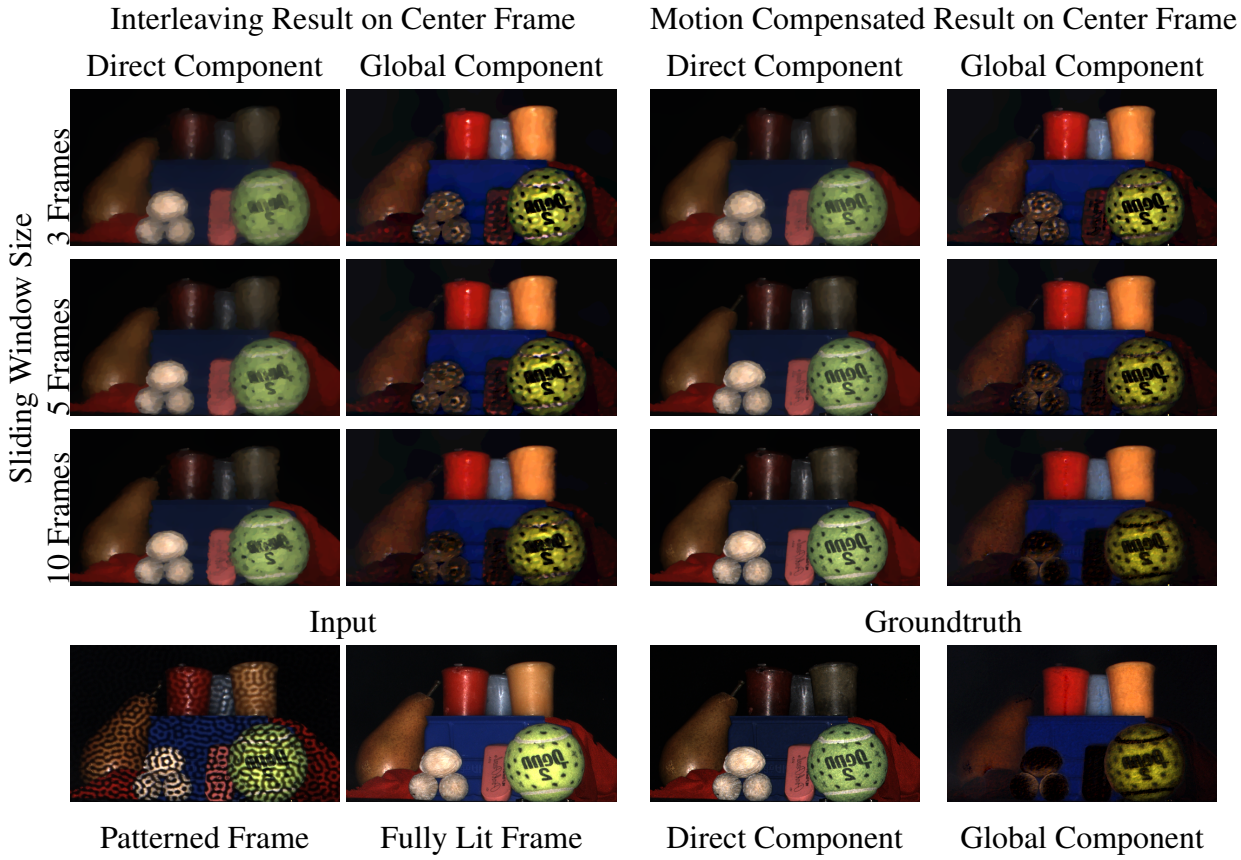


Figure 2.7: Comparison to Interleaving: The separation from interleaving and our algorithm as the number of frames in the sliding window is changed. The camera is panning across the scene. The first three rows show the results generated by both methods for various sliding window sizes (global images are shown at 2 times actual brightness). Our method makes more efficient use of images than interleaving because no frames are needed exclusively for tracking. Separations that resolve a given level of detail can be obtained with a smaller temporal sliding window than an interleaving approach. For instance, our method resolves the text on the tennis ball much earlier and more clearly than interleaving.

images. These aligned images are then separated as described in 2.3.3.

When the number of frames used is small, the regularized static method and the proposed motion compensated methods perform similarly. As the number of frames increases, the improvement in the motion compensated output reduces and then stops. When the window size is large, the frames near the edges of the sliding window can not be aligned to the center frame because the viewpoint changes are too large and the global and direct components of the scene points change appreciably. The motion compensation algorithm automatically discards these frames and they yield no improvement in the results.

Interleaving does not perform as well as our proposed method. The temporal window available for performing separation on dynamic scenes is small. With interleaving, only half the frames in this window can be used for computing separation. Additionally, if the motion is not smooth, the interpolated warps do not align images accurately. These problems could be solved by using higher frame rates, but at a given capture rate our method can handle faster, more complex motion than interleaving. Fig. 2.7 shows results from our motion compensation algorithm and interleaving with different temporal window sizes in an example scene.

2.4.3 Deformable Motions

In addition to rigid motion, our algorithm can also compensate for deformations and articulated motions. Figs. 2.8 and 2.9 show example separations obtained from videos of moving hands, deforming faces and a plant blowning in the wind. For comparison, we have included the results without motion compensation. When motion estimation is switched off, we select the window size for each example that gives the best result to compare against our method. Without motion compensation, the results are blurred and edges (around the fingers for example) are corrupted. Subsurface scattering occurs in the skin so most of the color appears in the global component.

2.5 Discussion

Although we do not model the changes in global and direct components that occur within a small temporal window, our method is still able to handle broad specular lobes like shiny surfaces on wax and highlights on skin. Sharp specularities and specular inter reflections such as those from polished metal surfaces would cause both the image alignment and component separation steps to break down. The fast direct-global separation algorithm for static scenes can handle sharp specularities but not specular inter reflections. One solution would be to use crossed polarization filters to remove specular reflections. Methods for removing specular components from images include [54] and [62].

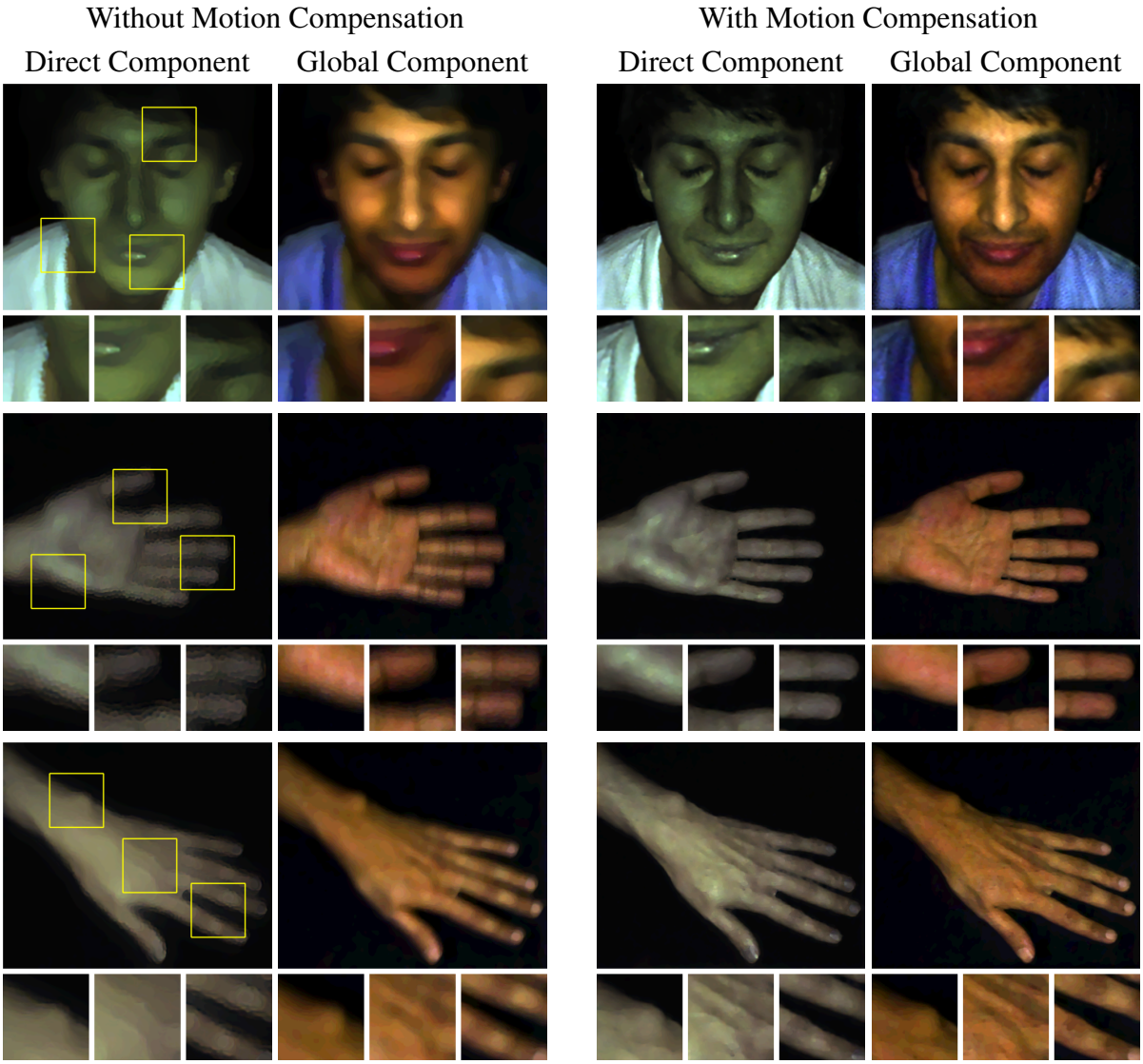


Figure 2.8: Direct-Global Separation on Skin: On the left are the global and direct components estimated without any motion compensation for a face changing expression and articulating hands. Many details get blurred away like the hair, and the lines on the palm. Other motion artifacts are clearly visible on the fingers. The two columns on the right show the component estimates on the same frames using our motion compensation method. With motion compensation, many of the artifacts are corrected and a lot more of the original scene detail is recovered. In the face example, some of the subsurface scattered light leaks into the direct images giving them a slightly greenish tinge. This happens because the subject is somewhat further away and the projected patterns at the subject are not high enough frequency to completely resolve the subsurface scattered light. A similar effect was observed in [70].

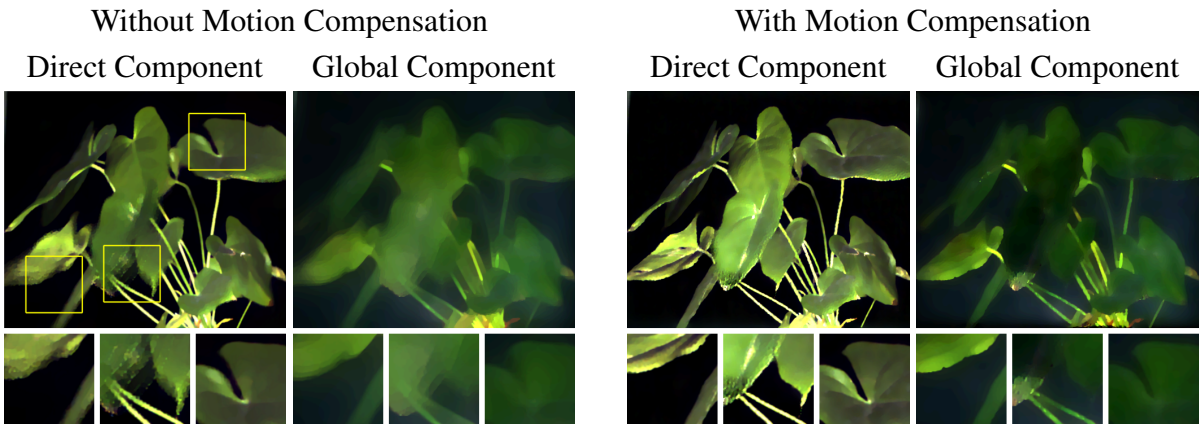


Figure 2.9: Direct-Global Separation on a Plant: On the left are the global and direct components estimated without any motion compensation for a plant moving in the breeze. There are motion artifacts and ghosting around the leaves. Using our motion compensation method, many of the artifacts are corrected and a lot more of the high-frequency details- like veins on the leaves are recovered.

The experiments presented were designed so that motion blur and defocus would not cause additional errors. In less controlled settings, we would need to consider the trade-off between acquisition time and accuracy. Using shorter exposure times and smaller apertures to avoid motion blur and defocus means that less light reaches the camera and image noise becomes more of a problem. We would need to consider how computational photography methods like coded aperture for motion deblurring [83] and light efficient photography [36] could be applied.

In chapter 4 we build a projector-camera based on a different projector technology and demonstrate a purely optical method for capturing different components of illumination. The method involves no computation and it can be applied to moving scenes. It can handle not only subsurface scattering and diffuse interreflections but also specular reflections and caustics.

Chapter 3

Multi-Focus Structured Light

"Most of the time, you'll be calibrating."

3.1 Introduction

Projector technologies like DLP and LCoS use lenses with large apertures to focus their output. The large aperture size is needed to maximize light throughput but as a result active illumination systems that use these types of projectors are limited to a shallow working volume in which the projector is in focus. This limits their applicability to scenarios where the scene relief is small and laboratory or industrial settings where the relative geometry between the objects of interest and the projector-camera system can be carefully controlled.

Pattern coding strategies like gray codes [46] degrade gracefully when illumination is defocused. In [26] patterns are designed such that they are all attenuated to roughly the same extent by projector blur and [42] uses a sliding projector as the light source. These methods have some robustness to illumination blur but they do not explicitly model illumination blur and use a single projector focus setting. When depth variation in a scene is very large, the structured light patterns in some areas will be blurred too severely for shape recovery to be possible.

In this chapter, we present a structured light algorithm that extends the working volume of the projector-camera system and is capable of producing high resolution depth maps over large working volumes. Our algorithm models both illumination defocus and global illumination effects like scattering and interreflection. In addition to a depth map of the scene, our algorithm recovers the direct and global components of illumination. It can be used to scan optically challenging materials like wax, marble and translucent plastic.

A naïve approach to expanding the depth of field would be to project a complete set of structured light patterns at each focus setting and then combine the resulting depth maps, but such an approach would require an inordinately large number of images. Our algorithm uses

multiple focus settings but projects only a small number of patterns at each setting, keeping the overall number of images required small. The key insight of our method is that even an illumination pattern that is not in focus at a scene point can aid in pattern decoding, provided the projector blur kernel has been carefully characterized. We perform this characterization by calibrating the projector to find the blur kernel as a function of scene point depth for each focus setting.

Previous work in structured light associates a fixed, depth independent code word with each projector pixel. In contrast, in our approach a projector pixel's code has a defocus induced dependency on the depth of the point it is illuminating. To test a candidate projector-camera pixel correspondence hypothesis, we first compute the scene point depth implied by the hypothesis. This depth value can be used to predict the defocused illumination received by the scene point from the projector. If the candidate correspondence is correct, this projector output prediction should match well with the intensity values observed at the camera pixel. By using a range of focus settings, we ensure that at least some segment of a projector code is always in sharp focus at a point in the scene. Our algorithm seamlessly combines two complementary depth cues - triangulation based cues which provide high depth resolution but require sharp illumination focus (and thus suffer from narrow working ranges) and defocus based cues which work over a large range of depths but provide coarse depth estimates. Our shape recovery algorithm is purely temporal and does not use spatial windows for decoding projector patterns which allows it to recover high quality depth maps with few artefacts at scene discontinuities. Once the shape has been predicted, we automatically have an estimate of the illumination received by each point of the scene in each image. We use this information to recover the direct and global components of illumination.

3.1.1 Related Work

The idea of exploiting projector defocus as a cue to recover shape was proposed in [110]. The approach involved estimating a measure of the projector pattern blur occurring at each illuminated scene point and mapping this measure to a depth value using a calibration function. They could recover accurate depth maps, but the fixed blur-to-depth mapping could not handle global light transport effects like sub-surface scattering. Gupta et al. [29] proposed a method to simultaneously model both projector defocus and global illumination. Their technique allows for depth recovery in the presence of global illumination and is based on the observation that unlike defocus blur, the blur induced by global light transport effects is almost independent of projector focus. Both [110] and [29] use colocated projector-camera systems and recover depth solely from focus/defocus cues.

In contrast, our approach does not use a colocated configuration but performs stereo triangu-

lation between the camera and projector to measure depth. It has been shown that in principle, depth from defocus is similar to stereo triangulation [88] but focus/defocus cues have a baseline equal to the size of the aperture. Since triangulation cues are computed over the wider projector-camera baseline, our method is capable of producing more fine grained depth estimates. Although we do not use defocus cues explicitly (by using an illumination sharpness measure for instance), they are used implicitly as our projector codes are modeled as being depth dependent due to defocus. Previous work that combines camera defocus and stereo cues includes [109] and [103].

In structured light literature, many methods proposed to prevent errors due to global light transport rely on high spatial frequency illumination patterns. Examples include [12, 14] and [26]. In [27], global illumination effects are handled by designing a set of light pattern codes that work well with long range effects like inter reflections and a second set of patterns that work well with short range effects like sub-surface scattering. For scenes with both types of effects, ensembles of codes are generated and a voting scheme is used to estimate depth. Unlike [27], we do not seek to assign a binary code to each pixel and instead attempt to fit a model to the observed projector and camera values at a pixel, so we can use a single set of patterns to handle both types of global illumination effects.

Micro phase shifting [26] is a phase shifting variant that uses a narrow band set of high frequency sinusoids as the projected patterns. Because the patterns all have similar frequency they are attenuated similarly by projector defocus which lends some robustness to projector blurring. However, it should be noted that while this has some robustness to blur, it does not model defocus or use multiple focus settings so it can not handle large variations in scene depth.

In [60] illumination defocus is exploited towards a different end. Sinusoidal patterns are generated by projecting binary patterns with a defocused projector. DLP projectors can project binary patterns at very high frame rates which allows the phase shift algorithm to run in real time and recover dynamic scenes.

3.2 Modeling Image Formation and Illumination

Let $S^t(x)$ be the value of the projected structured light pattern at time t at a scene point imaged by camera pixel x. The brightness $I^t(x)$ observed by a camera pixel is a weighted sum of the direct illumination $I_d(x)$ and the global illumination $I_g(x)$ of the scene point. When the pattern $S^t(x)$ has a high spatial frequency and a 50% duty cycle, it can be shown that the contribution of the global illumination to the observed brightness is approximately pattern independent and equal to $\frac{1}{2}I_g(x)$ [69]. The pattern modulates the direct component so its contribution to the observed brightness is $S^t(x)I_d(x)$. Thus we have

$$I^t(x) = \tfrac{1}{2} I_g(x) + S^t(x) I_d(x) \tag{3.1}$$

We had used the same image formation model earlier in chapter 2. We use π to denote the correspondence between projector pixels and camera pixels that illuminate/image the same scene point, $p = \pi(x)$. The projector value seen at time t at a scene point at depth z illuminated by projector pixel p, is a defocused version of the projector pattern value at that pixel $L^t(p)$. It has been shown that unlike camera defocus blur, the defocus blur kernel for a projector is scene independent in the sense that the kernel at a scene point depends only on the depth of the point, not on the geometry of the neighborhood surrounding the point [110]. Thus, without resorting to assumptions like local fronto-planarity, the effects of projector defocus blur can be modeled by convolving the projector pattern $L^t(p)$ with a spatially varying blur kernel $G(p, z, f)$.

$$S^t(x) = \tilde{L}^t\left(\pi\left(x\right)\right) = \left(L^t * G\left(\pi\left(x\right), z, f\right)\right)\left(\pi\left(x\right)\right) \tag{3.2}$$

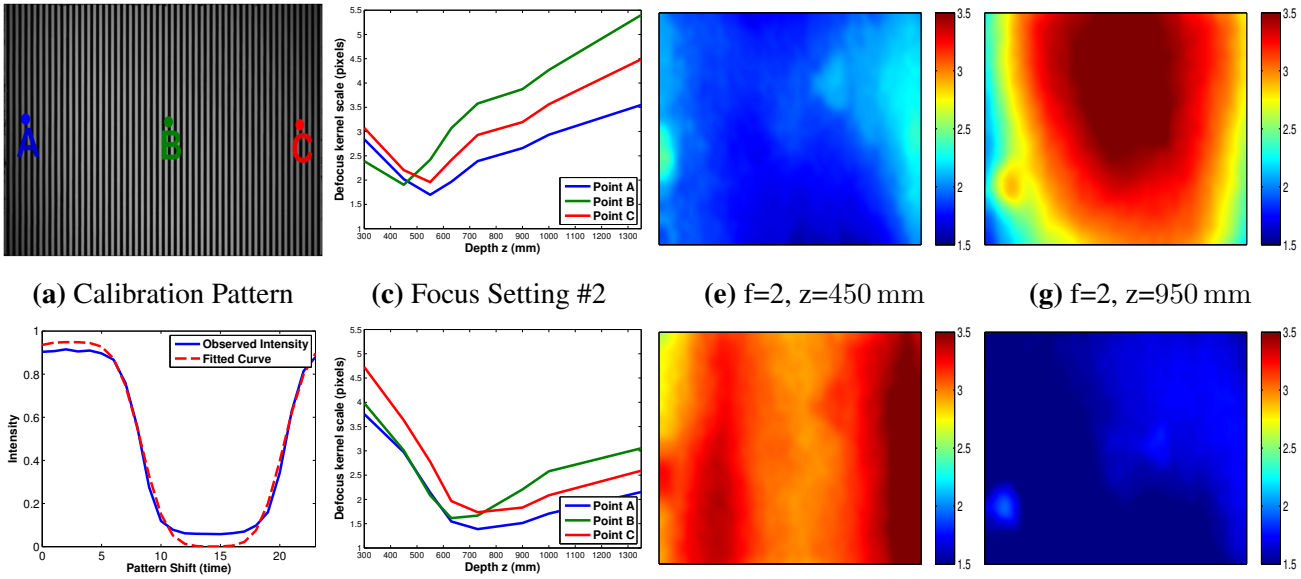
The blur kernel G depends on the scene point depth z and the projector focus setting f. Additionally, we allow the function G to vary spatially with projector pixel coordinate as this helps better model the projector's optical aberrations.

Although the original high frequency illumination pattern $L^t(p)$ is blurred due to defocus, Equation 3.1 still holds. The defocus blur reduces the amplitude of the high frequency components of the pattern but does introduce any low frequency content into the signal. We use a small aperture on the camera ($f/10$ in our experiments) and model it as a pinhole camera that does not introduce any additional blurring due to camera defocus.

Characterizing Projector Defocus

We model the projector blur using a spatially varying, isotropic Gaussian kernel. The scale of the blur kernel $\sigma(p, z, f)$ is a function of projector pixel location p, the depth z of the scene point being illuminated and the current focus setting of the projector f. A more general class of kernels may allow for a more accurate characterization and allow more complex types of aberrations to be modeled, but we found that isotropic Gaussians were sufficient for our purpose.

For a given focus setting f and target depth z we estimate the defocus blur by projecting a sequence of patterns onto a planar target at depth z. The patterns are horizontal square waves with a period of 24 pixels (fig. 3.1a). We capture 24 images as the pattern translates one pixel at a time. The temporal profile of intensity values observed at a pixel is modeled as a square wave convolved by the blur kernel (fig. 3.1b). A similar scheme was used in [110] to estimate a mapping between illumination defocus and scene point depth. We find the blur kernel scale $\sigma(p, z, f)$ that best fits the observed temporal profile for each projector pixel. This characterizes



(a) Calibration Pattern **(c)** Focus Setting #2 **(e)** f=2, z=450 mm **(g)** f=2, z=950 mm

(b) Blur Kernel Fit **(d)** Focus Setting #5 **(f)** f=5, z=450 mm **(h)** f=5, z=950 mm

Figure 3.1: Characterizing Projector Defocus: (a) - image of one of the square wave patterns for estimating projector blur. (b) is the temporal intensity profile at point B and the Gaussian smoothed square wave fit. (c) and (d) - the blur kernel scale σ for the projector pixels A,B and C for two different focus settings as the scene depth is varied. (e) to (h) - maps of blur scale σ across the projector image for different combinations of focus setting and scene point depth. The value of σ clearly varies across the image, especially when the projector is out of focus.

the defocus blur at one depth and focus setting (example σ maps are figs. 3.1e-1h). We repeat the process at a set of depths for each focus setting ($f = 1, 2, ...|F|$). We sample $G(p, z, f)$ at every projector pixel p and focus setting f, but only sparsely in depth z. When queried for the blur kernel at a given focus setting and depth, we return the kernel at that focus for the nearest calibrated depth. Projector characterization is a one time, off line process.

3.3 Illumination Control and Image Acquisition

We recover shape and perform direct-global separation with a set of structured light patterns captured at different projector focus settings. The focus settings are chosen so that the projector's plane of focus spans the entire working volume of the scene and that every part of the scene has at least one setting where the illumination is in reasonably good focus. For each of the F focus settings we capture a small number (N) of structured light patterns. Although we have chosen to capture an equal number of patterns at each setting, this is not a requirement for the algorithm,

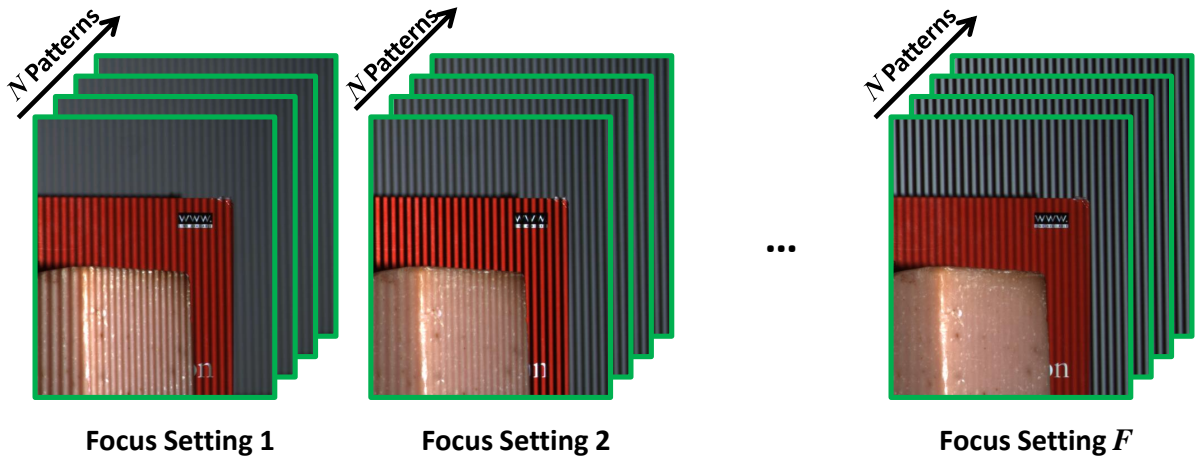


Figure 3.2: Input to our Algorithm: We use binary stripe patterns of varying width. Unlike most other structured light algorithms that use a fixed focus setting on the projector, we change the focus setting to move the plane of focus backwards during the image capture process (the camera focus however remains fixed). We capture a total of $T = F \times N$ images. In our experiments, T typically ranged between 20 and 30. As the figure illustrates, near by objects receive focused illumination in the earlier parts of the sequence and distant objects come into focus later on.

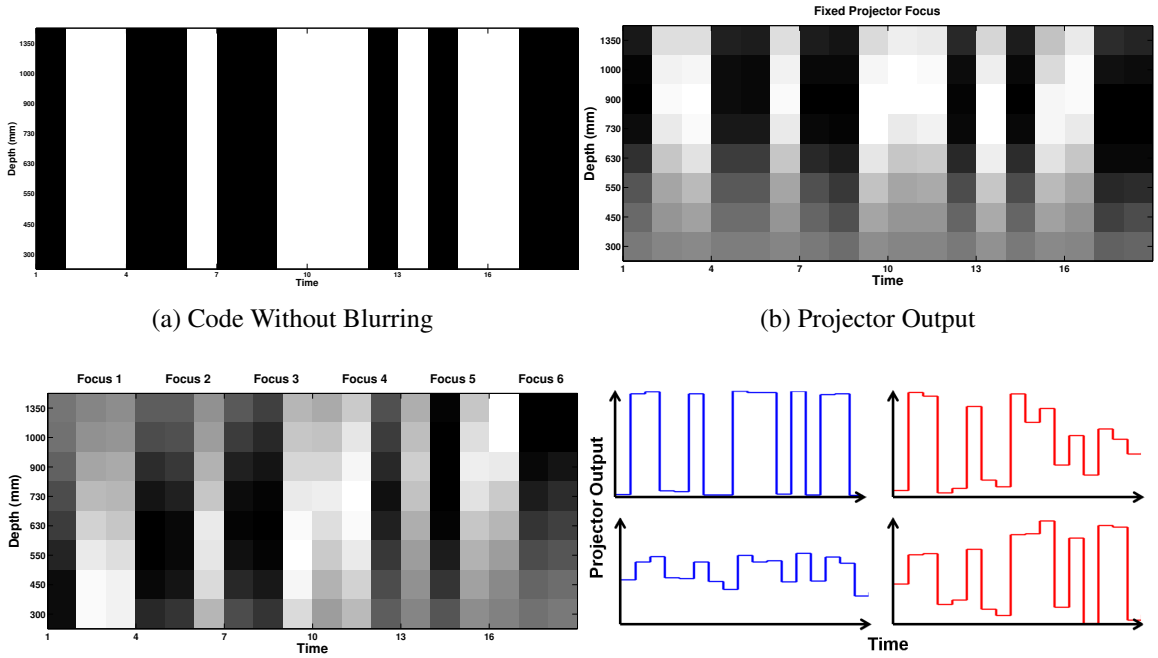
the number of patterns used could be varied adaptively depending on the scene.

The structured light patterns we use are vertical binary stripes with randomly varying widths. Higher frequencies are less susceptible to global illumination errors, but very high frequency patterns are not displayed well by projectors. We let the period of the stripes in a pattern fluctuate between 10 and 14 pixels. This frequency range is high enough to prevent global illumination errors in most situations while still being in the band where contemporary projectors works effectively. We select patterns that do not correlate with each other to ensure that there is little redundancy between patterns.

3.4 Recovering Shape With Defocused Light Patterns

Temporal structured light algorithms project a series of patterns onto the scene, the time sequence of values emitted by a projector pixel form a code for that pixel. Camera-projector correspondence is established by finding the projector code that best matches the time sequence of intensity values observed at each camera pixel. The code can be binary (eg. gray codes) or continuous (eg. phase shifting), but it assumed that the code for each projector pixel is independent of the scene geometry.

In contrast, our multi-focal structured light algorithm explicitly models illumination defocus



(a) Code Without Blurring

(b) Projector Output

(c) Our Multi Focus Code

(d) Comparison of Codes

Figure 3.3: Effect of Defocus on Codes: When illumination defocus is not modeled, the temporal code associated with a projector pixel (a horizontal cross section of (a)) is independent of depth. However, as (b) shows, outside a narrow working range, the actual appearance of the code is depth dependent. In (c), we use 6 focus settings and 3 patterns per focus. Using multiple focus settings allows us to expand the systems working volume. Also, we model illumination defocus so blurred codes do not cause errors. When regular codes are in focus, they work well (upper blue graph in (d)), however for scene points that are out of focus, contrast is very poor (lower blue graph in (d)). On the other hand, our multiple focus codes always have parts that are well focused and thus have high contrast (red graphs in (d)).

effects, so a projector pixel's code becomes a function of the depth of the scene point it is illuminating. This idea is illustrated in figure 3.3. It is clear, that when the depth variation in a scene is large, defocus can strongly affect how a projector code manifests in a scene. As seen in figure 3.3b, when a pattern is out of focus, different values become difficult to distinguish. Decoding such a blurred pattern reliably with a defocus-blind algorithm would necessitate very high illumination power and high dynamic range on the imaging sensor. As figure 3.3c shows, even in large working volumes, some part of our code is always in sharp focus. This allows our method to work at lower illumination power levels over extended depths.

If we hypothesize that projector pixel p corresponds to camera pixel x, we can perform triangulation to find the scene point depth $\tau_z(x, p)$ implied by the hypothesis. Using our defocus

model (equation 3.2), we can then simulate the projector value $\tilde{S}^t(x,p)$ that would be observed at this scene point by convolving the projector illumination pattern L^t with the defocus kernel,

$$\tilde{S}^t(x,p) = \left(L^t * G\left(p, \tau_z\left(x,p\right), f\right)\right)(p) \tag{3.3}$$

Stacking together these values for all the patterns $t = 1, ..., T$ gives us the projector code for the pixel.

$$\tilde{S}(x,p) = [\tilde{S^1}(x,p), \tilde{S^2}(x,p), ..., \tilde{S^T}(x,p)] \tag{3.4}$$

This projector code needs to be matched against the sequence of observed intensity at camera pixel $I(x)$

$$I(x) = [I^1(x), I^2(x), ..., I^T(x)] \tag{3.5}$$

If the hypothesis that pixel x and pixel p correspond to each other is correct, then by our illumination model (equation 3.1), there should be a linear relationship between the observed intensity values at the camera and the simulated projector values. We quantify the quality of a camera-projector correspondence hypothesis by computing the correlation coefficient between $I(x)$ and $\tilde{S}(x,p)$. We can then find the projector pixel $p = \pi(x)$ corresponding to camera pixel x by maximizing this correlation.

$$\pi(x) = \underset{p}{\operatorname{argmax}}\ \rho\left(I\left(x\right), \tilde{S}\left(x,p\right)\right) \tag{3.6}$$

We use a calibrated projector-camera system so with the epipolar constraint we limit the search in equation 3.6 to a $1D$ search along the epipolar line. We compute $\rho(I(x), \tilde{S}(x,p))$ for every p along the epipolar line corresponding to a positive depth (Figure 3.4). To compute disparity to sub-pixel accuracy, we interpolate ρ scores between projector pixels when searching for maximae.

3.5 Recovering Direct and Global Illumination Components

Once the camera-projector correspondence map π has been estimated, we can compute $S^t(x)$, the projector pattern value at each camera pixel taking defocus blur into account using equation 3.2. Under the image formation model (equation 3.1), there is a linear relationship between the projected pattern value at a point $S^t(x)$ and the brightness observed by the camera $I^t(x)$. Fitting a line to this model at each pixel yields estimates of the global and direct illumination. However, it is possible that even over the entire sequence of projected light patterns, some camera pixels would have seen seen only a small range of projector intensity values. There will be significant

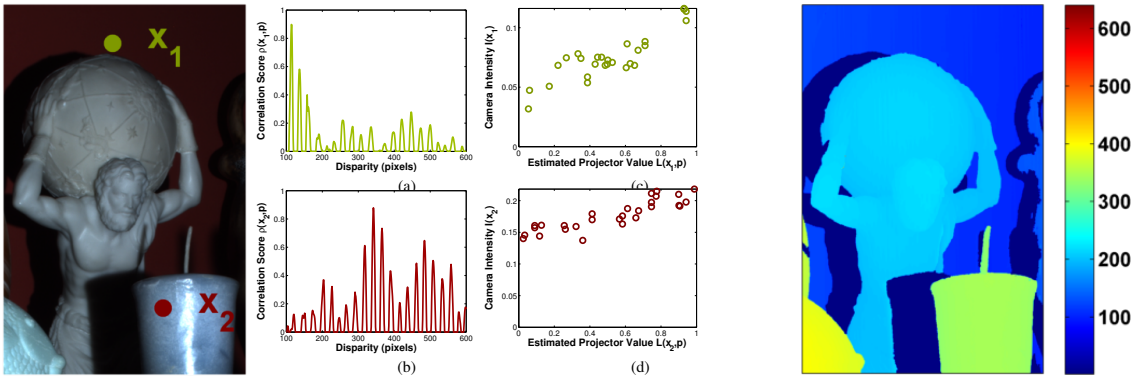


Figure 3.4: Establishing Correspondences: Part of a scene (left) and the computed disparity map (right). Graph (a) shows the correlation score for point x_1 as a function of disparity to the projector. The disparity that leads to the best match is 115. There are many peaks in the correlation score graph, but modeling of illumination blur causes the peaks to decay as we move away from the correct disparity value. Graph (c) shows the intensity observed by the camera against the simulated projector illumination value for the best disparity. Graph (b) and (d) are the same trends for point x_2. Because of strong sub-surface scattering at x_2, the global illumination component is large and the direct component is relatively small. This can be seen in (d).

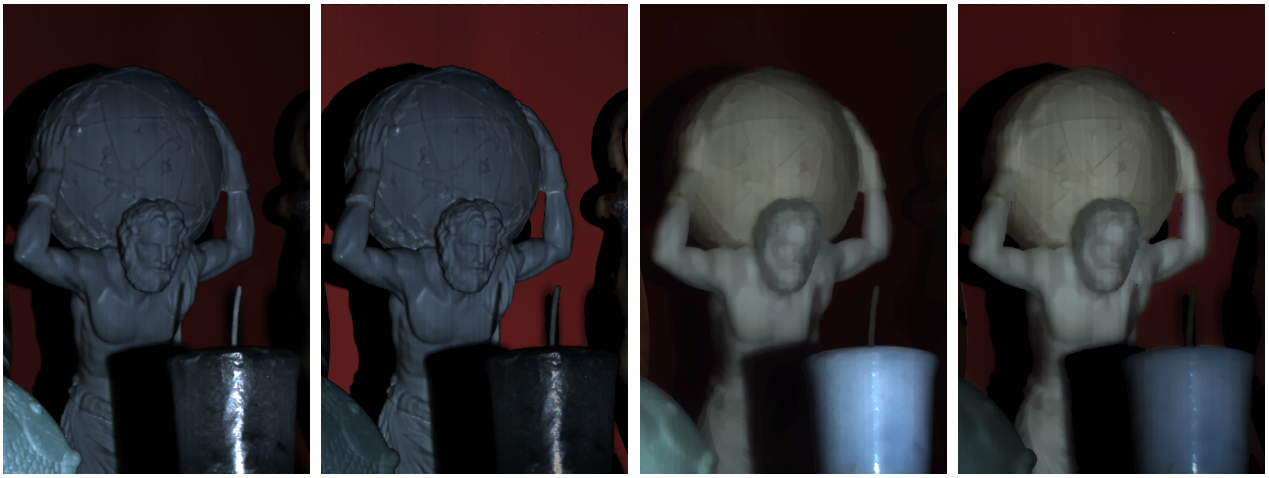
ambiguity while fitting a line to data at these pixels and hence there will be numerous plausible solutions to the direct-global separation. We resolve these ambiguities using a smoothness prior as was done in chapter 2 by finding the direct image I_d and global image I_g that solve

$$\underset{I_d, I_g}{\operatorname{argmin}} \sum_{t \in T} \|I^t - \tfrac{1}{2} I_g - S \circ^t I_d\|_2^2 + \lambda_d TV(I_d) + \lambda_g TV(I_g) \tag{3.7}$$

λ_d and λ_g are scalar parameters that weight the smoothness terms for the direct and global components respectively. $A \circ B$ is the Hadamard (element-wise) product between A and B. $TV(F)$ is the isotropic total variation of the function $F(x, y)$

$$TV(F) = \sum_{Domain(F)} \sqrt{\left(\frac{\partial F}{\partial x}\right)^2 + \left(\frac{\partial F}{\partial y}\right)^2} \tag{3.8}$$

Parts of the scene far away from the projector receive less light that regions close to the projector. As a result, there is a pronounced falloff in the recovered direct and global illumination images. Because we have recovered scene geometry, we can roughly correct for this falloff by assuming it follows an inverse square relationship with depth. We can compute depth dependent correction factor $K(x)$ at each pixel



(a) I_d (b) $\bar{I}_d$ (c) I_g (d) $\bar{I}_g$

Figure 3.5: Direct Global Separation: The direct (a) and global (c) components of illumination estimated by our algorithm. Due to camera saturation on the specular highlights, part of the specular reflection from the candle surface leaks into the recovered global image. Since we have recovered a depth map of the scene, we can also correct for projector fall off. This is particularly useful in scenes with large depth variations where objects in the background appear much darker than those in the foreground because they are further away from the light source. After accounting for the fall off, we get corrected estimates for the direct and global component illumination (images (b) and (d) respectively).

$$K(x) = \frac{\alpha}{\tau_z^2(x, \pi(x))} \tag{3.9}$$

where α is an (arbitrary) positive scale factor. We can then solve for the corrected direct and global illumination components ($\bar{I}_d$ and $\bar{I}_g$) by modifying equation 3.7:

$$\underset{\bar{I}_d, \bar{I}_g}{\operatorname{argmin}} \sum_{t \in T} \|I^t - \tfrac{1}{2} K \circ \bar{I}_g - K \circ S \circ^t \bar{I}_d\|_2^2 + \lambda_d TV(\bar{I}_d) + \lambda_g TV(\bar{I}_g) \tag{3.10}$$

3.6 Results

Experimental Setup

Our experimental setup consists of a projector and a camera mounted in a stereo configuration. We use a 500 lumen DLP projector with a resolution of 1280×800 (InFocus IN1144). The camera is a 2448×2048 color CCD (Point Gray Research GRAS-50S5-C). The camera is calibrated geometrically and radiometrically. The projector is calibrated geometrically at each focus setting

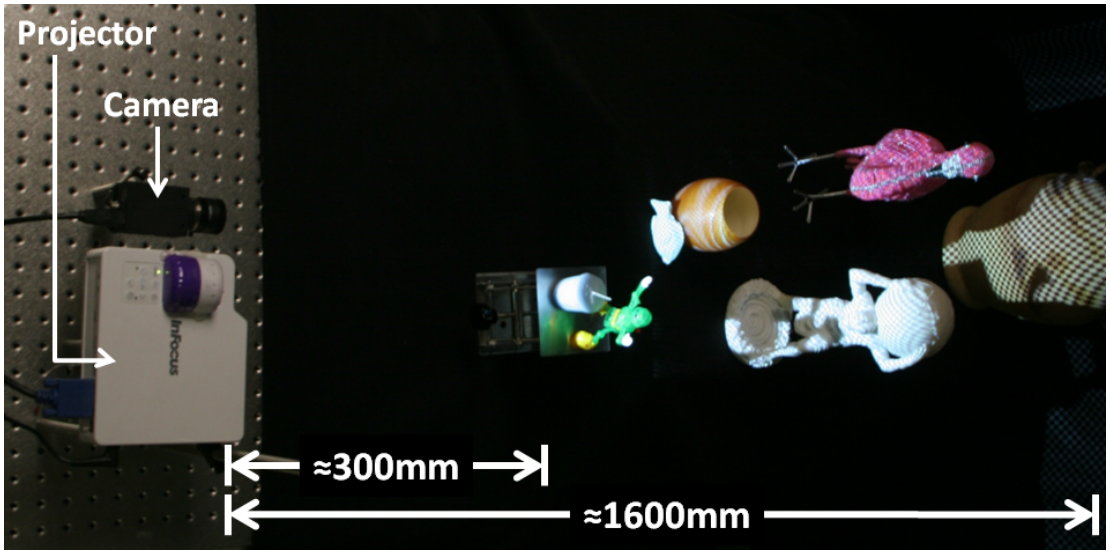


Figure 3.6: Experimental Setup: With a small, 500 lumen DLP projector that has limited depth-of-field and a small number of images, we are able to scan scenes over a large working volume to recover accurate depth maps and perform illumination separation.

and its blur kernel has been characterized (as described in Section 3.2). Our method uses only binary stripe patterns but we calibrated the projector radiometrically so that we could compare our method to micro phase shifting [26]. The projector-camera baseline is fixed and known. Since the projector intrinsics change with focus setting, we correct for this by warping images before projecting them so that geometrically they all appear to be projected by a projector with fixed intrinsic parameters.

In our experiments the focus ring position was changed by hand and we used 4 positions ($F = 4$). Between the shortest and longest focus settings, the working range of the system covers depths from 350mm to 1600mm. For all experiments the camera lens aperture was set to $f/10$, the exposure time was 133ms and the camera was configured to capture 8 bit images.

3.6.1 Depth Recovery

We present results from our depth map recovery algorithm on two challenging scenes (top row in fig 3.7). Depth maps from our algorithm (second row in fig 3.7) were generated using 7 structured light patterns at each of 4 focal lengths, a total of 28 images. Our algorithm is able to recover accurate depth maps of both scenes with very few errors. We compare against a simple depth from illumination focus algorithm (bottom row) and micro phase shifting (third row).

The illumination defocus algorithm we compared against projects a shifted sequence of square waves (14 images) at each of 8 projector focus settings and then finds the focus setting

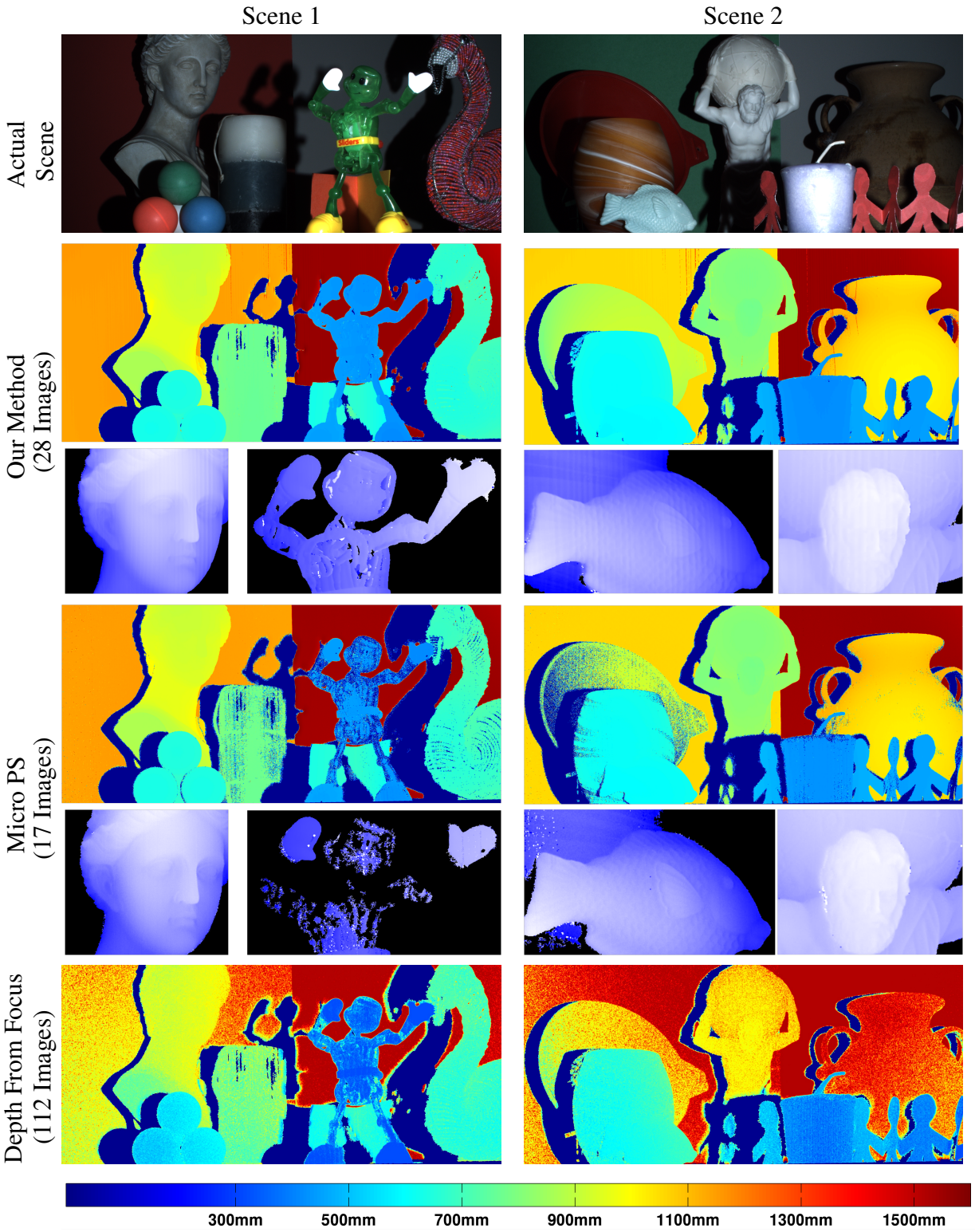


Figure 3.7: Our algorithm recovers depth maps for scenes containing challenging objects over an extended working volume with relatively few images. Insets show (rescaled) depth maps for small parts of the scene. Fine details like the scales on the soap fish are successfully resolved.

at which each camera pixel's illumination contrast was maximized. Each focus setting can be mapped to the depth of its corresponding plane of focus to find the depth map. Since the baseline for this method is limited to the aperture of the projector, the resulting depth estimates are coarse along z and tend to be inaccurate at large distances.

For the micro phase shifting experiments, we chose the high frequency (16 pixels per cycle) pattern set with 15 frequencies [26]. Micro phase shifting uses a fixed projector focus so we set the projector to be in focus in the middle of the scenes. The total number of patterns used is 17. Using more patterns at this frequency is difficult because micro phase shifting requires all projected pattern frequencies to be in a narrow band.

Micro phase shift has some robustness to illumination blur but since it does not actually model defocus, it breaks down when the depth variation in a scene is too large. This is evident in scene 1 where the shape of green plastic robot in the foreground is not recovered by micro phase shifting. In comparison, our method is able to recover the robot. Our algorithm also works better on low albedo or poorly lit regions like the red funnel in scene 2 because there are guaranteed to be some images in which these objects are well focused. Since we change focus settings, there are always some images where the contrast of our projected illumination is high, so low signal to noise ratios are less of a problem for our algorithm.

The candle in scene 1 is very difficult to reconstruct as from some directions, it reflects almost no light directly back to the camera, almost all the observed radiance is due to sub-surface scattering. As a result, all the methods are unable to recover depth at some points on the candle surface.

3.6.2 Recovering Direct and Global Illumination

To obtain ground truth direct and global illumination images for our scenes, we projected 14 shifted stripe patterns at 8 projector focus settings and used the multiple focus separation technique proposed in [29]. The results presented for our method are computed using the same 28 images that were used to estimate the depth maps. Although our technique uses fewer images and involves a smoothing term, it generates output that is similar to the ground truth. Additionally, we can correct for the effects of projector fall off as demonstrated in Figure 3.9.

3.7 Discussion

We presented an algorithm that can use a projector-camera system to reconstruct shape and recover direct and global illumination in a large working volume with a small number of images, despite the fact that projector's have small depths-of-field. The naive approach of collecting a full

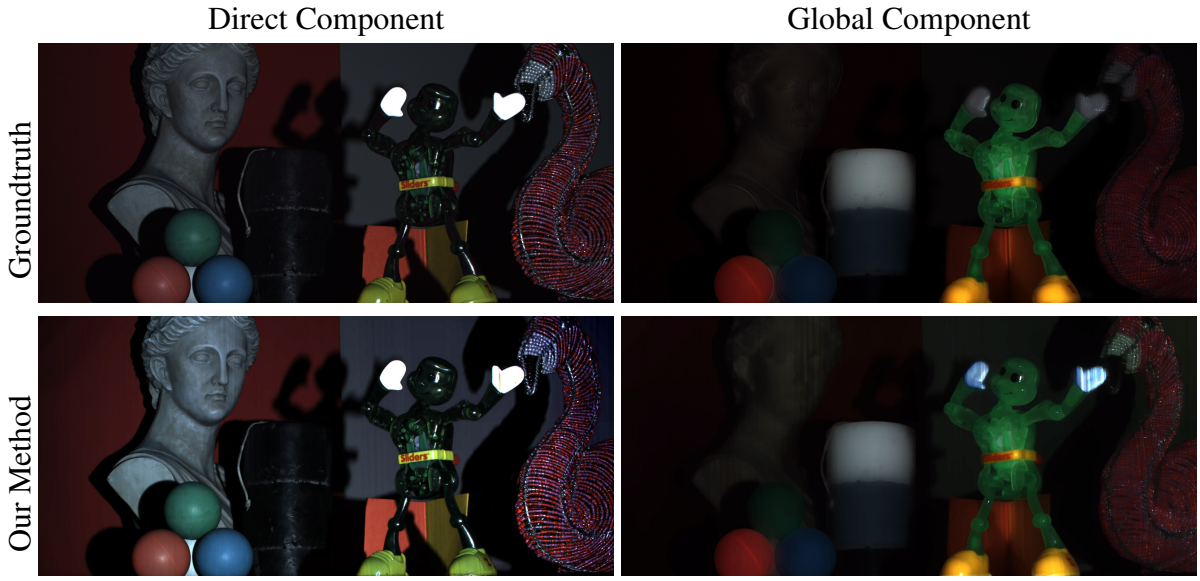


Figure 3.8: Direct-Global Separation. Groundtruth was computed using 112 images and our method used 28. The white hands on the robot toy appear much brighter in the global image computed by our method than the ground truth. This is because our algorithm tried to fit a linear trend to saturated (completely white) camera pixels.

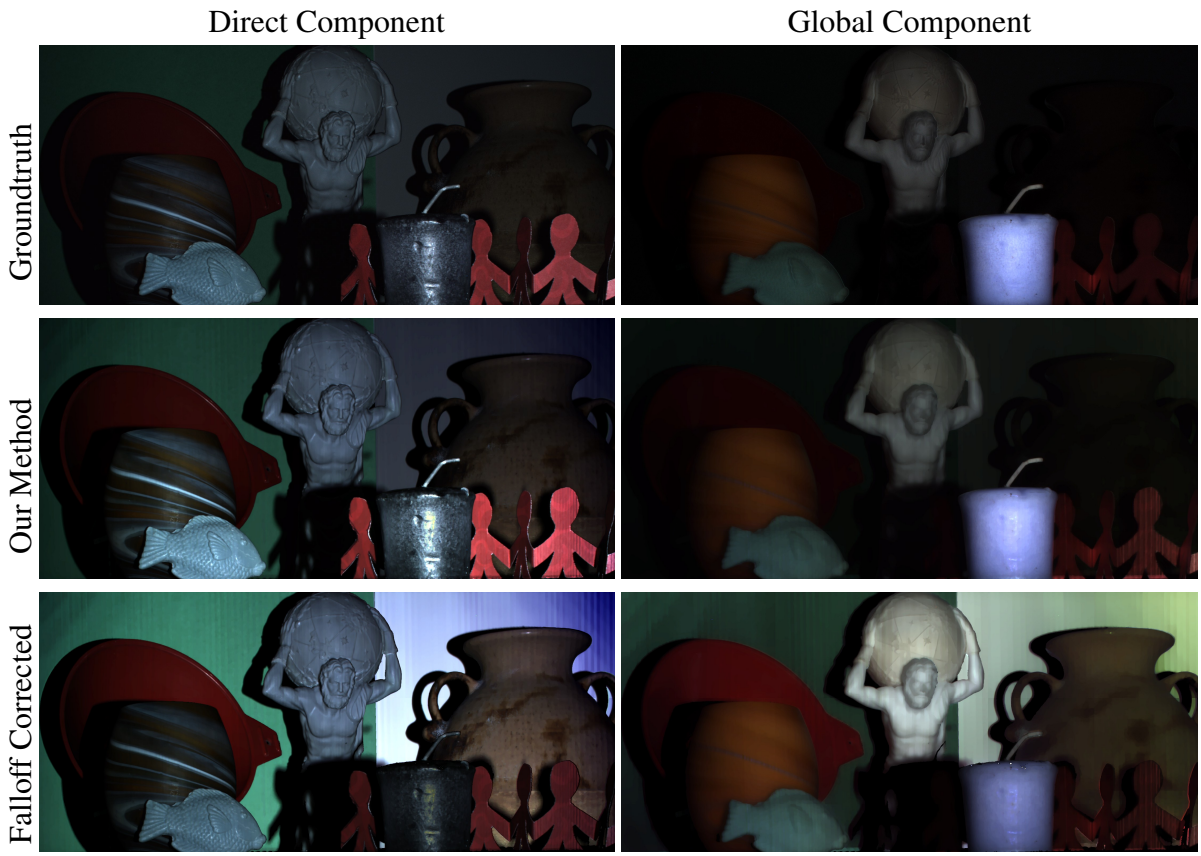


Figure 3.9: Direct-Global Separation. Groundtruth was computed using 112 images and our method used 28. In this scene, the shading on the soap fish and the white statue becomes is very clear in the direct illumination image. As our method also simultaneously recovers depth, we can correct the direct and global images for the the projector fall-off with depth.

set of structured light images at each of many focal settings and then combining the results would require a very long acquisition time. The key insight of our method is that even an out-of-focus image contains useful information if the blur characteristics of the projector are known.

Our algorithm's robustness to global illumination effects relies on the assumption used in [69]- the global illumination must vary slowly compared to the spatial frequency of the projected patterns. If this assumption does not hold, for example when specular interreflections occur, our method fails.

We currently use randomly chosen stripe patterns. Optimal pattern sets for structured light are usually derived by trying to maximize the distance between codes to minimize the chance of a decoding error. In our setting, we would have to consider the fact that defocus causes our codes to vary with depth. Also, for the direct-global component separation to work well, each pixel's code word must contain a large range of projector intensity values. Carefully designed patterns may allow our algorithm to work well with fewer images.

We currently do not model the effects of camera defocus blur. In our experiments, we minimize camera defocus by using a small aperture $f/10$ while the projector lens has a very large aperture and hence a narrow depth-of-field. Projector defocus is much easier to model than camera defocus because the blurring occurs in the projector's image plane as opposed to in the world [110]. However, for a given total amount of blur across the projector-camera system light throughput is maximized when the projector and camera apertures are matched. Extending our model to handle camera blur could potentially help increase light throughput and decrease net acquisition time.

Chapter 4

Episcan3D

"The best way to solve a problem is to not have one."

4.1 Introduction

In this chapter, we describe the Episcan3D sensor - an energy-efficient, portable active illumination device that can measure the shape in scenes with strong global transport and can handle very high levels of ambient illumination despite using a low power projector as the light source and regular CMOS cameras for imaging. Unlike chapters 2 and 3 which used regular DMD or LCoS-based projectors, in this chapter we use a redisitributive projector as the light source. The robustness to global illumination and ambient light that the proposed device achieves stems from the fact that that it can selectively image light paths and uses a hardware configuration that illuminates and images a single line in the scene at a time. This line-by-line scanning approach strikes a balance between point-by-point scanning (which is robust, but slow) and strategies that illuminate and image the scene in its entirety at once (which are fast, but sensitive to ambient light and global light transport). The sensor can recover shape using structured light or active stereo and can also probe components of light transport optically at video frame rates.

4.1.1 Epipolar and Non Epipolar Probing

In chapters 2 and 3 we relied on high spatial frequency illumination patterns to handle the effects of global illumination. In this chapter, we use an alternate approach - we design the imaging system that so it captures only the components of light transport that we are interested in.

Selective imaging of light paths (probing light transport) can be performed using a projector light source and a camera with a programmable mask [75]. The direct component of illumination always obeys the epipolar geometry constraint between the light source and sensor. In contrast,

for most scenes, only a very small fraction of global light transport falls along the epipolar lines. This observation was the key insight of [76] where it was used to design probing operators that made structured light scanning robust to the effects of global light transport.

The Episcan3D sensor relies on this same concept of probing and the geometric intuition concerning projector-camera light transport. It can efficiently capture images containing only the epipolar component of illumination (which is the direct illumination plus the small fraction of global illumination that obeys epipolar geometry) or it can capture images containing only the non-epipolar component of illumination (which contains only global illumination). Both these probing operations are implemented purely optically, without the need for image subtraction or any other operation in software to be performed on the raw output from the camera.

4.2 Energy Efficient Probing

By epipolar geometry, the camera pixel corresponding to a projector pixel is constrained to lie along a line in the camera image. When the projector and camera are physically arranged so that they are in a rectified stereo configuration, each row of projector pixels (we will henceforth refer to projector rows as scanlines) corresponds to a row of pixels in the camera image.

Conceptually, implementing epipolar-only (or non epipolar-only) probing should be simple. Illuminate a single projector scanline at a time and expose only the corresponding row of camera pixels (for non epipolar-only imaging, mask only the corresponding row). Repeat this process for each scanline and then read out the image from the camera sensor. However, probing using this pattern-mask sequence with a regular projector that can not redistribute its light output (such as those based on DLP or LCoS technology), only block or unblock individual pixels is extremely inefficient in terms of light throughput. Using a regular projector with m scanlines, if only one scanline is turned on while the others are off, only $1/m^{\text{th}}$ of the light output of the projector is emitted into the scene. This means that the light throughput of the probing operation will only be $1/m^{\text{th}}$ (or 0.14% for a projector with 720 scanlines). For the task of non epipolar-only probing, there exist pattern and mask sequences based on Hadamard codes have a light throughput of 25% [76], but for epipolar-only imaging there is no light efficient solution with a non-redistributive projector.

This efficiency problem can be solved by using a raster scanning laser projector. A scanning projector uses a laser light source and a 2D MEMS mirror. The mirror deflects the laser beam along a raster scanning path while the laser turns on and off to project the desired pattern onto the scene. A scanning projector can be thought of as an impulse projector that redistributes all of its output energy into one point at a time or at a coarser timescale, into a single scanline at a time. Now, implementing epipolar-only probing by illuminating one scanline at a time and exposing

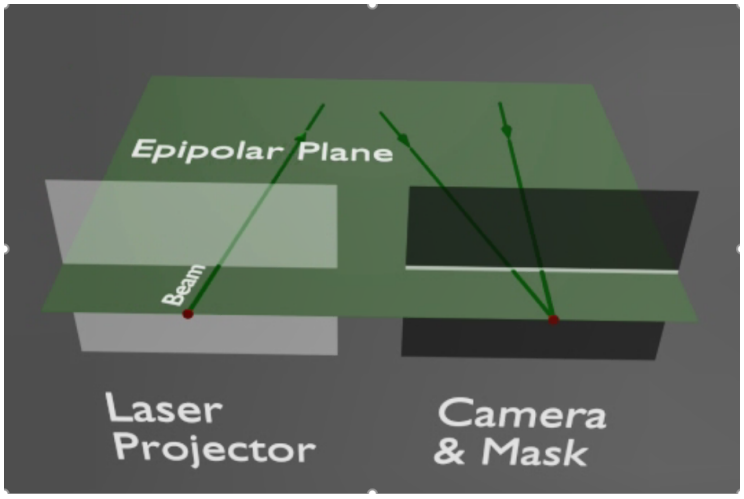


Figure 4.1: Energy-Efficient Epipolar Probing. A raster-scanning projector is combined with a camera sensor with maskable pixels. The projector and camera are placed in a rectified stereo configuration so that by epipolar geometry, each projector scanline corresponds to a row of pixels in the camera. To perform epipolar-only probing the camera mask is configured so that at any point of time, only the row of pixels corresponding to the active projector scanline is exposed. To implement non-epipolar probing, the camera masks are inverted - only the row corresponding to the active scanline is masked, all other pixels are exposed.

only the corresponding rows of camera pixels is a very efficient operation (see figure 4.1). In fact, this scheme is optimal in terms of light throughput and probing accuracy because

- no light is blocked coming out of the projector
- no light traveling along epipolar paths is blocked at the camera sensor
- no light traveling along non-epipolar paths reaches the camera sensor.

Optimal non epipolar-only probing can be performed by inverting the camera side masks. For a more thorough analysis of light throughput and methods for deriving optimal pattern/mask sequences to implement probing operations using different types of redistributive light sources refer [77].

An important property of epipolar-only imaging with a scanning laser source is that each pixel is exposed only for a very short time. For a projector with m scanlines, each pixel is exposed only $1/m^{\text{th}}$ of the time. During this short pixel exposure, all the light traveling along epipolar paths between the projector and camera pixel will be captured but very little ambient light will be integrated by the sensor. The apparent intensity of ambient light sources in the scene is reduced by a factor of m. This allows our sensor to work effectively outdoors despite using a very low power projector as its light source.

Another interesting property of epipolar-only imaging is that it partially masks out the effects of defocus blur on the camera side. Only the parts of the blur point spread function that lie along the epipolar line will contribute to the captured image.

Epipolar imaging also mitigates the effects of intereference between multiple active-illumination devices operating simultaneously in the same environment. Consider the two device case: each device would illuminate and image a single line in the scene at a time, so at any instant the second device can only interfere with the first at the points where its illuminated line intersects with the first system's exposed row of pixels. Over the course of an image capture, this point of interference would trace out a curve through the image. Instead of interefering everywhere the fields of view overlap, the two devices would interfere with each other only along these curves. If two devices were to align perfectly and just happened to be synchronized with each other, they would potentially interefere strongly everywhere, but this would be an extremely unlikely occurence.

4.3 Prototype Hardware

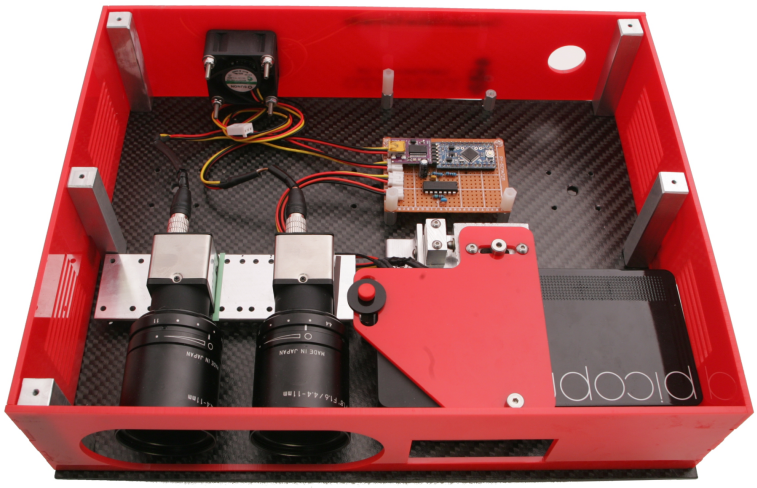


Figure 4.2: Episcan3D Hardware Implementation: The Episcan3D sensor consists of a low power raster scanning laser projector, a pair of CMOS rolling shutter cameras and a microcontroller for synchronization. There is no dedicated hardware for masking on the camera side as the camera rolling shutters are used to perform masking.

The current hardware implementation of the Episcan3D sensor is shown in figure 4.2. The sensor consists of a projector light source, two cameras and a microcontroller to synchronize the cameras to the projector. Only one camera is needed to implement probing operations, having two cameras running simultaneously in epipolar-only imaging mode allows us to implement

energy efficient active stereo as we shall discuss later.

The light source is a raster scanning laser projector (PicoPro from Celluon Inc.) with a resolution of 1280×720. The projector has three laser light sources - red ($639\,\mathrm{nm}$), green ($522\,\mathrm{nm}$) and blue ($445\,\mathrm{nm}$). The total power output of the projector across all three channels is 30 Lumens. The projector has a refresh rate of $60\,\mathrm{Hz}$ and takes approximately $20\,\mu\mathrm{s}$ to draw a single scanline. Drawing all 720 scanlines takes $14.4\,\mathrm{ms}$, during the remaining time in the $16.67\,\mathrm{ms}$ projector cycle, the projector mirror returns to the home position to start the next raster scan. The projector's horizontal field of view is approximately $45°$.

The cameras used are monochrome CMOS rolling shutter cameras (UI-3250CP-M-GL Rev.2 from IDS Imaging Development Systems GmbH). The cameras have $1/1.8''$ sensors and a resolution of 1600×1200. The cameras run at up to 60fps. The cameras use low distortion varifocal lenses (Lensagon CVM0411ND from Lensation GmbH). Also, each camera is fitted with a wavelength filter having a $25\,\mathrm{nm}$ passband and $645\,\mathrm{nm}$ center frequency to match the red laser output of the projector (645HBP25 from Omega Optical Inc.).

The cameras and projector are on custom machined mounts that allow for their positions to be adjusted so that they are correctly rectified. The projector has been modified to generate a VSYNC signal at the start of each frame. There is no dedicated hardware to implement masking on the camera side.

4.3.1 Camera-side Masking With Rolling Shutters

For epipolar-only and non epipolar-only imaging with a laser projector and a rectified camera, the camera masks take the form of lines. For epipolar-only imaging, the entire sensor is masked except for one row of pixels at a time and for non epipolar-only imaging these masks are inverted. It is possible to implement these types of masks very effectively in a rolling shutter camera without the need for additional hardware.

Figure 4.3 shows our prototype's timing diagram. We use t_p to denote the duration for which the projector dwells on a single scanline. The speed at which the rolling shutter progresses down the rows of the image (t_c) is determined by the pixel clock, we can choose the pixel clock and focal length on camera lens so that the rate at which the exposed camera row moves down the image matches the rate at which the projector scanline changes. Increasing the camera exposure (t_e) increases the thickness of the band of camera pixels exposed for each projector row. Changing the delay (t_o) between the VSYNC signal from the projector and the trigger signal passed to the camera changes the offset between the illuminated row on the projector and the imaged row(s) on the camera.

To capture the epipolar component, the exposure t_e for each camera row is matched to the time the projector stays on a scanline (t_p) and the other timing parameters are chosen so that the

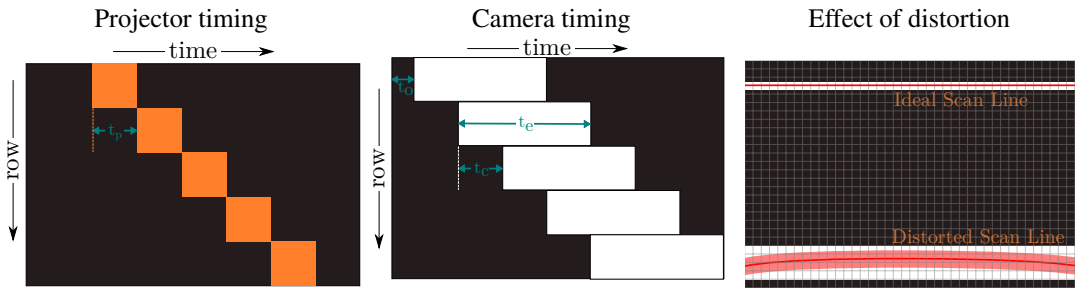


Figure 4.3: Electronic masking with a rolling shutter: At each timestep (of duration t_p), the projector illuminates a single scanline (top left). The camera uses its rolling shutter to selectively mask pixels (top right). The masks are defined by three controllable parameters, the exposure time t_e, the time it takes the rolling shutter to read a row of pixels t_c and the offset between the projector sync output and the camera trigger input t_o. Ideally, a single projector scanline corresponds to a single camera row. In practice, because of distortion and jitter, each scanline corresponds to a band of camera pixels.

line scanned by the projector is synchronized to the row being exposed in the camera. Conversely, to capture non-epipolar light, the camera exposure time is set to be t_p less than the projector cycle time and the trigger is offset by t_p so that every row is exposed for the entire projector cycle except during the time it is illuminated directly by the projector.

Ideally, we would be able to configure the rolling shutter so that only the rows of camera pixels illuminated by the projector at any timestep would be exposed (Figure 4.3). In practice, the projector we use generates distorted scanlines that are not absolutely straight. Additionally, we observe synchronization jitter and small perturbations in the trajectory of the projector's laser during each exposure. This means that the region in the camera image corresponding to a projector scanline is constrained to lie inside a narrow band in the image, not along a line. To accommodate these bands we thicken the region of unblocked pixels in each mask by increasing the pixel exposure time t_e and adjusting the trigger offset t_o. With our current system we had to set the t_e to 100μs during epipolar-only imaging to counter the effects of jitter and scanline distortion. In the ideal case, we would have been able to set t_e equal to t_p (around 20μs). As a result, some ambient light leaks into the epipolar only image. Also, some short-range indirect light, like parts of the sub-surface scattered component leak into the epipolar-only image instead of being part of the indirect-only image.

In computational photography, camera sensor masking is often implemented by coupling a DMD or LCoS spatial light modulator to the sensor with a relay lens. This approach was adopted in [84, 102] for compressive image sensing and in [76] for probing light transport. These arrangements are more flexible than rolling shutter masking, but they require many extra optical

components. These components increase system complexity, introduce aberrations, severely reduce optical efficiency and reduce working volume.

4.4 Results

We first demonstrate the ability of the Episcan3D sensor to capture the epipolar-only and non epipolar-only components of light transport efficiently at video frame rates. We then show how epipolar-only imaging can be used to make structured light scanning robust to the effects of global light transport and ambient light. By operating two cameras instead of one in epipolar-only imaging mode we can implement energy-efficient active stereo, suitable for shape recovery in dynamic scenes. We then make a small modification to the configuration of the projector and camera that allows us to optically section scenes and selectively capture light paths based on the disparity.

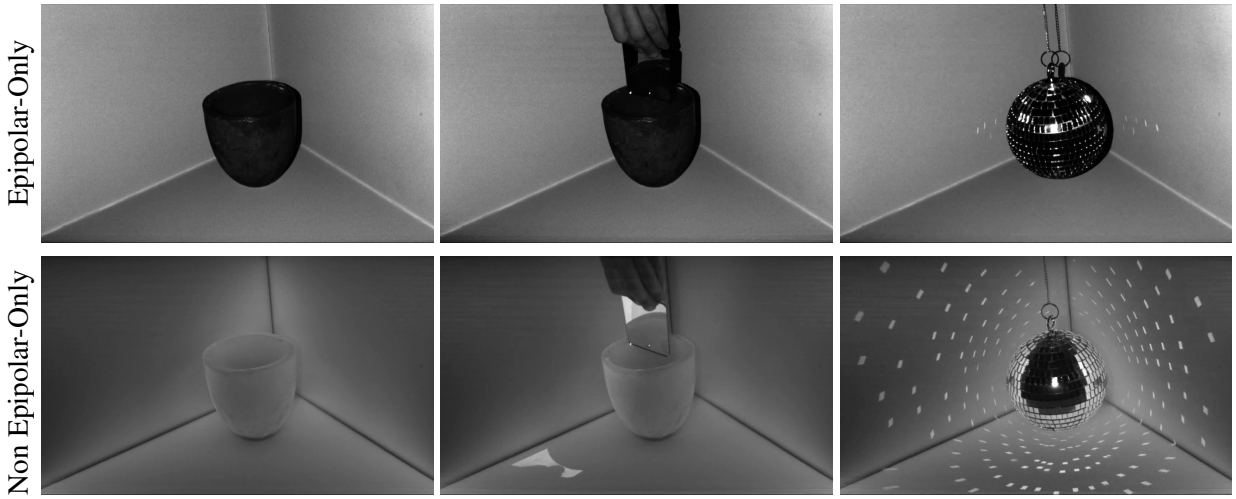


Figure 4.4: Live Transport Probing: Epipolar-only imaging captures direct light transport and the small fraction of global light transport that obeys epipolar geometry. Non Epipolar-only probing captures only global light. These are example frames taken from a video where the sensor alternated between epipolar-only imaging and non epipolar-only imaging. Diffuse interreflections between the walls appear in the non-epipolar image. The wax bowl appears dark in the epipolar-only image because most of its apparent brightness is due to subsurface scattering. Most specular interreflections from the mirror (except those that happen to obey epipolar geometry) appear in the non epipolar-only image.

Some of the results presented below were obtained using an older version of the Episcan3D sensor. The working principle of the older version is identical, but it used a lower resolution

(800×480) laser projector (ShowWx+ from MicroVision Inc.) with a lower output power (15 Lumens). The camera used was also different, a 1280×1024 color CMOS camera (UI-3240-C-HQ from IDS Imaging Development Systems GmbH).

4.4.1 Live Transport Probing

Figure 4.4 shows images captured during epipolar-only and non epipolar-only probing with our sensor. Subsurface scattering, diffuse interreflections and specular interreflections all appear predominantly in the non-epipolar only image because light paths corresponding to these phenomena do not obey epipolar geometry. Due to radial distortion of the projector scanlines and synchronization jitter, the sensor has to expose a small band of camera pixel rows at a time instead of a single row. Hence some non-epipolar light that is close to following epipolar geometry will appear in the epipolar-only image.

4.4.2 Structured Light Scanning of Difficult Objects

Epipolar-only imaging can be used in conjunction with any structured lighting technique to provide robustness to global light transport. The left side of figure 4.5 shows the result of structured light shape recovery with normal imaging on a set of challenging objects. The structured light patterns used are a high frequency variant of gray codes and are designed to be robust to global light transport effects (see "Pattern Design" below for details). Despite the fact that high frequency patterns provide some robustness to global light transport, all the reconstructions contain errors due to scattering and/or interreflections.

With the same patterns and decoding algorithm, epipolar-only imaging yields much more accurate results as seen on the right side of figure 4.5. Diffuse and specular interreflections are suppressed and contrast loss do to subsurface scattering is avoided.

Pattern Design In structured light, some robustness to global light transport effects can be obtained by using structured light patterns that do not contain low spatial frequencies. The binary-reflected gray code sequences are the most commonly used binary patterns for structured light scanning. Gray codes use low spatial frequencies for the more significant bits of the codes, similarly phase shift profilometry typically uses low frequency patterns for phase unwrapping / disambiguation. As a result, neither of these techniques are robust to global light transport. Micro Phase Shifting [26] uses only sinusoidal patterns selected from a narrow band of high frequencies. XOR codes [27] are a gray code variant that avoid the use of low frequency patterns. The raster scanning projector used by Episcan3D was unable to accurately generate the sinusoidal patterns used by Micro Phase Shifting and XOR codes are sensitive to camera defocus

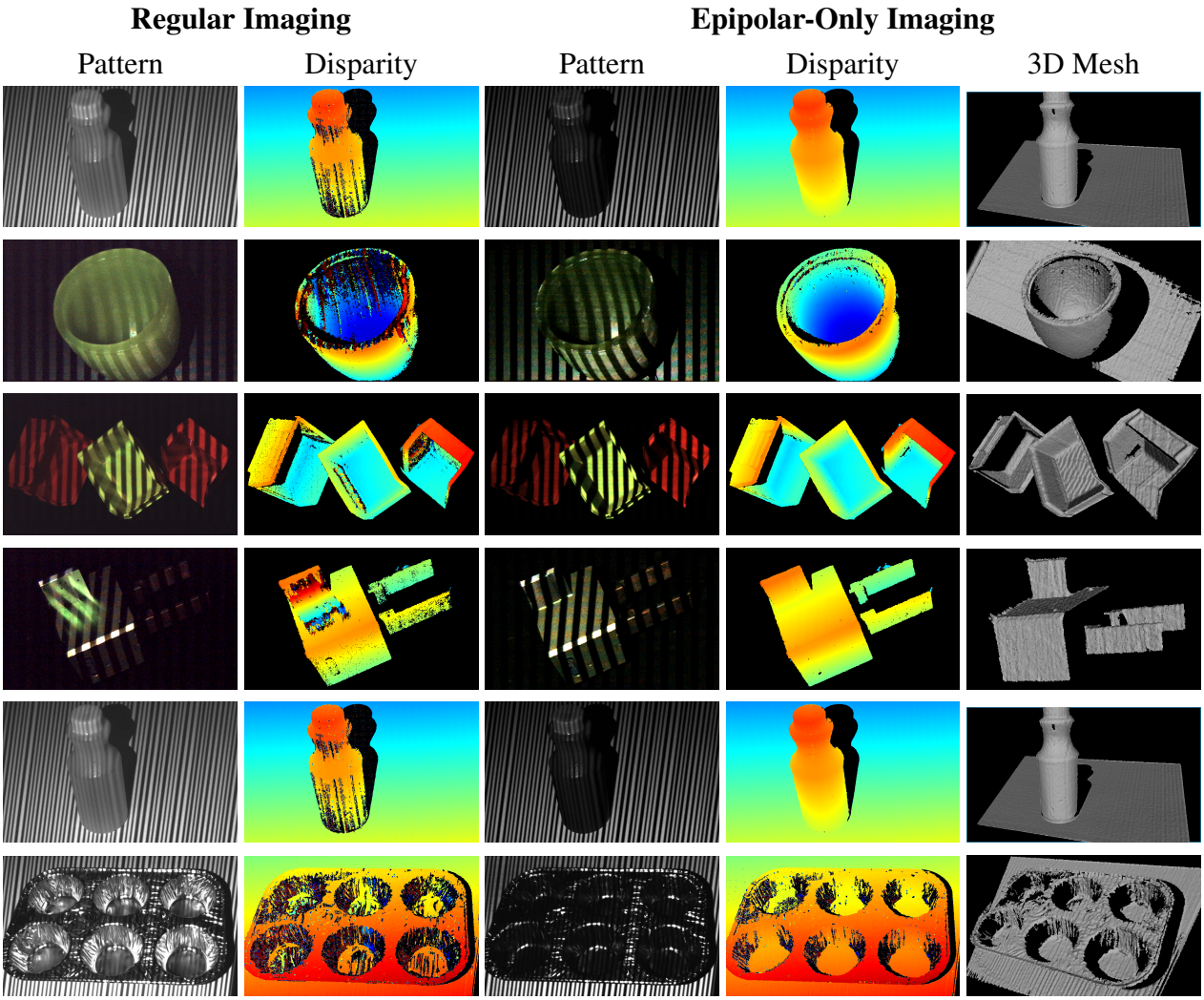


Figure 4.5: Imaging and reconstructing challenging objects with strong global illumination effects. Structured light with regular imaging is prone to errors due to global light transport effects (notice the effects of interreflections on the plastic bins, metal gripper and baking tray. Also observe the subsurface scattering in the green wax bowl). Epipolar-only imaging blocks almost all these effects. Notice how the cross hatching due to interreflections on the gripper has been eliminated and the pattern contrast on the bottle and wax bowl has increased. With the milk bottle, subsurface scattering causes contrast loss and blurring of the projected patterns with regular imaging, with epipolar imaging the patterns appear much sharper and as a result the depth estimates are more accurate. The baking tray is a particularly challenging example. Most of the errors due to specular interreflections have been suppressed, but some regions where interreflections fall along epipolar lines are still reconstructed incorrectly.

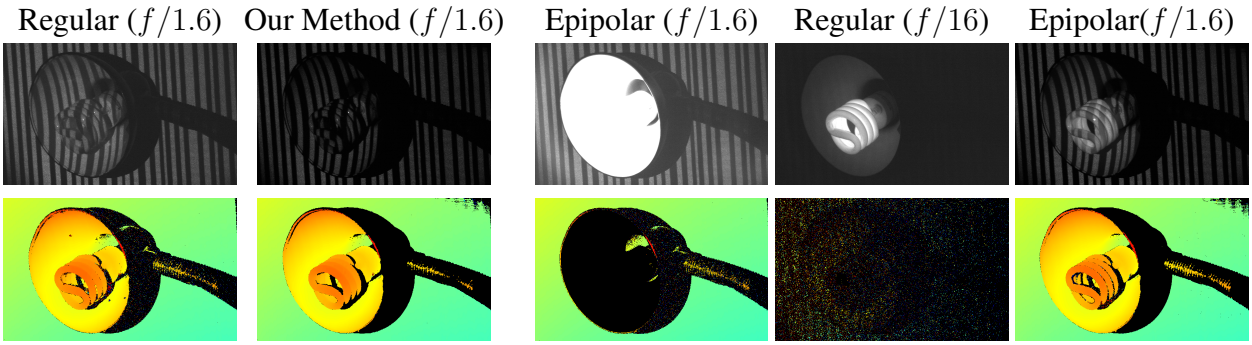
blur and prone to binarization errors. Instead, we designed a set of patterns based on gray codes that avoids the use of low frequency patterns (and is thus robust to global light transport) by solving the following constraints:

- Gray code property: Adjacent projector pixels should have binary codes that differ only by 1 bit (this leads to robust decoding at camera pixels that correspond to a mixture of two neighboring projector pixels).
- Minimum stripe width: The width in pixels of every high or low stripe in each projected pattern must be equal to or greater than the minimum stripe width (projectors typically perform poorly when projecting extremely high frequency patterns, such patterns are also very susceptible to defocus blur).
- Maximum stripe width: The width in pixels of every high or low stripe in each projected pattern must be equal to or less than the maximum stripe width (decoding errors due to global light transport effects are more likely with low spatial frequency patterns).
- Code Balance: The binary representation of each projector pixel's code shouldn't contain too many low bits or too many high bits (this ensures that the projected pattern at every camera pixel changes during the structured light pattern sequence, which allows the binarization threshold for each pixel can be determined automatically. The alternative is to binarize by projecting each pattern and its complement which doubles the number of images that need to be captured).

Schemes for finding gray codes with desired stripe width properties (also known as run lengths) are well established [23]. These would satisfy the first 3 criteria enumerated above, but not the last one (code balance). Such codings are robust to defocus and various global light transport effects and have been used for structured light depth recovery [34]. To find a suitable code that satisfies all four properties listed, we used a depth first search strategy. A set of 11 patterns are used (each projector row contains 1280 pixels so 11 binary patterns is the minimum number needed for decoding). We set the minimum stripe width to 6 pixels, the maximum stripe width to 15 pixels and the code balance parameter to 3 (i.e. each code word must have at least 3 high bits and 3 low bits in its binary representation).

4.4.3 Structured Light Scanning Under Bright Ambient Light

The epipolar-only imaging mode used by the Episcan3D sensor is robust to ambient light. Each row of pixels is exposed for a very short duration of time so little ambient light reaches the sensor. This is important because cameras have limited dynamic range and a bright ambient light source can easily overwhelm low power active light sources. We demonstrate this by projecting patterns onto a light bulb when it is turned on (figure 4.6), performing structured light scanning outdoors



(a) Lamp Off **(b)** Lamp On

Figure 4.6: Imaging and scanning a 1600 lumen lamp with a 15 lumen projector: When the lamp is off (a), the pattern projected by the projector is visible with both regular imaging and with epipolar-only imaging. By projecting a series of structured light patterns the lamp can be reconstructed. When the bulb is turned on (b), regular imaging breaks down. With a large aperture ($f/1.6$), the image is saturated, with a small aperture ($f/16$) the projected pattern is not visible to the camera on the bulb or the shade. In epipolar-only imaging mode, the Episcan3D sensor exposes camera pixels only while they can receive light from the projector and as a result, most of the light from the bulb is blocked, the pattern is clearly visible even on the bulb and the fixture can be reconstructed even though it is on.

on a bright, sunny day (figure 4.7) and scanning the objects used in section 4.4.2 under indoor spot lighting (figure 4.8).

4.4.4 Energy-Efficient Active Stereo

Passive stereo is widely used method for depth sensing. Establishing correspondence between images in a stereo pair is only possible in regions with strong image texture so simple stereo techniques often yield sparse depth maps. Non-local stereo algorithms can generate better results by integrating weak depth cues over large spatial regions and regularizing depth maps with various smoothness priors. These methods tend to be very computationally expensive.

A solution to this problem is to project a texture onto the scene. The projected texture makes it simple to compute depth in all parts of the scene within the projector working range, textured or untextured. Epipolar-only imaging makes it possible to project a pattern from a low power source that is visible to the camera even under bright ambient lighting. This is demonstrated in figure 4.9.

(a) Regular Imaging (High f-number)

(b) Epipolar-only Imaging (Low f-number)

(c) Disparity

(d) 3D Model

Figure 4.7: Active illumination in bright sunlight ($80\,\mathrm{klx}$): With regular imaging (a), the active illumination patterns are overwhelmed by sunlight and are not visible despite using a wavelength filter. Energy efficient epipolar-only imaging blocks a large fraction of the ambient light. This allows the projected pattern to be seen by the camera (b) and makes 3D structured light reconstruction possible (c,d).

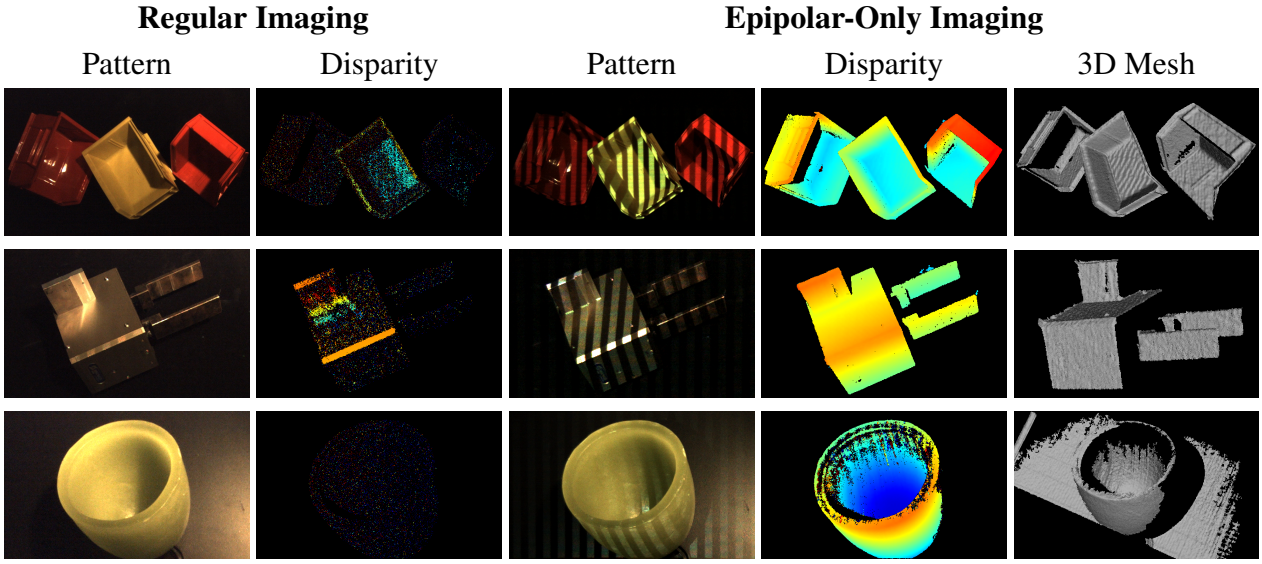


Figure 4.8: Scanning Objects Under Bright Ambient Lighting: These objects are being scanned using structured light while simultaneously illuminated by a bright indoor spotlight (ambient light level is approximately $10\,\mathrm{klx}$). In these examples, there is no wavelength filter attached to the camera. With regular imaging the patterns can not be detected by the camera and the reconstructions are poor. With epipolar-only imaging, the plastic bins and gripper are reconstructed well. The wax bowl is reflects very little light directly, most light scatters below the surface so it is particularly difficult to scan under bright ambient light.

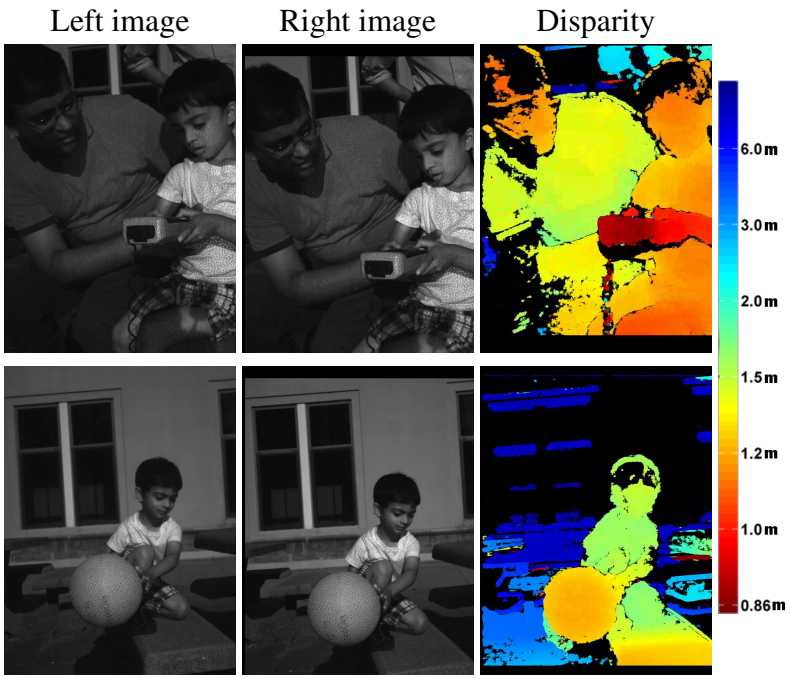


Figure 4.9: Energy Efficient Active Stereo: By operating a stereo pair in epipolar-only imaging mode we can image a projected texture from a low power projector outdoors on a bright sunny day ($85\,\mathrm{klx}$ ambient light level). This allows us to use a simple block matching stereo algorithm to generate dense stereo depth maps even on textureless regions of the scene as long as they are within the projector's working range (around $1.5\,\mathrm{m}$)

4.4.5 Disparity Gating

When the camera and projector are rotated by 90° so that projector scanlines and camera rows are perpendicular to the baseline, the correspondence between scanlines and rows becomes dependent on scene depth (see figure 4.10). By changing the trigger offset t_o we can optically mask light on the basis of scene point depth. By capturing a sequence of images where each image unmasks scene points at a narrow range of disparity values, we can recover the shape of a scene. In Figure 4.11 we demonstrate how disparity gating could be used to recover depth through a participating media.

4.5 Discussion

Epipolar-only imaging allows our sensor to work under bright ambient light conditions. We analyze how the sensor's effective working range is affected by ambient light levels and estimate the working range that could be achieved with a carefully engineered system based on our current

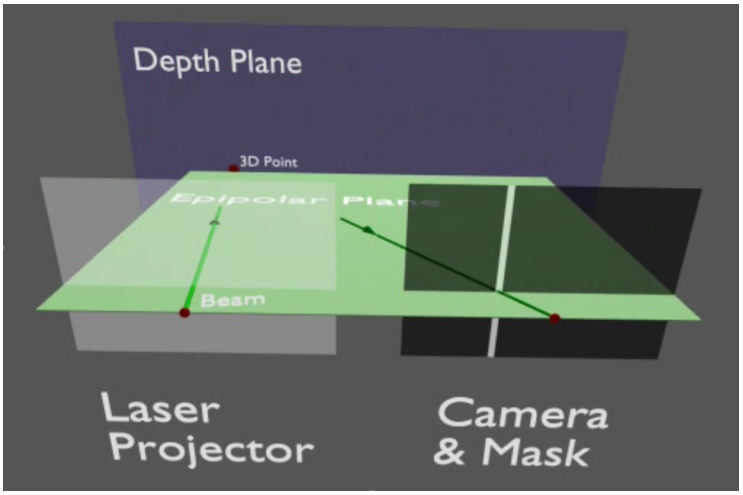


Figure 4.10: Geometric Arrangement for Disparity Gating. By rotating both the projector and camera in figure 4.1 by 90° around their optical axes, the projector scanlines and camera rows become perpendicular to the projector-camera baseline. If the projector and camera are synchronized so that there is a fixed offset between the active projector scanline and the exposed camera row, the image captured contains only light paths corresponding to a single disparity value - a single depth plane in the scene.

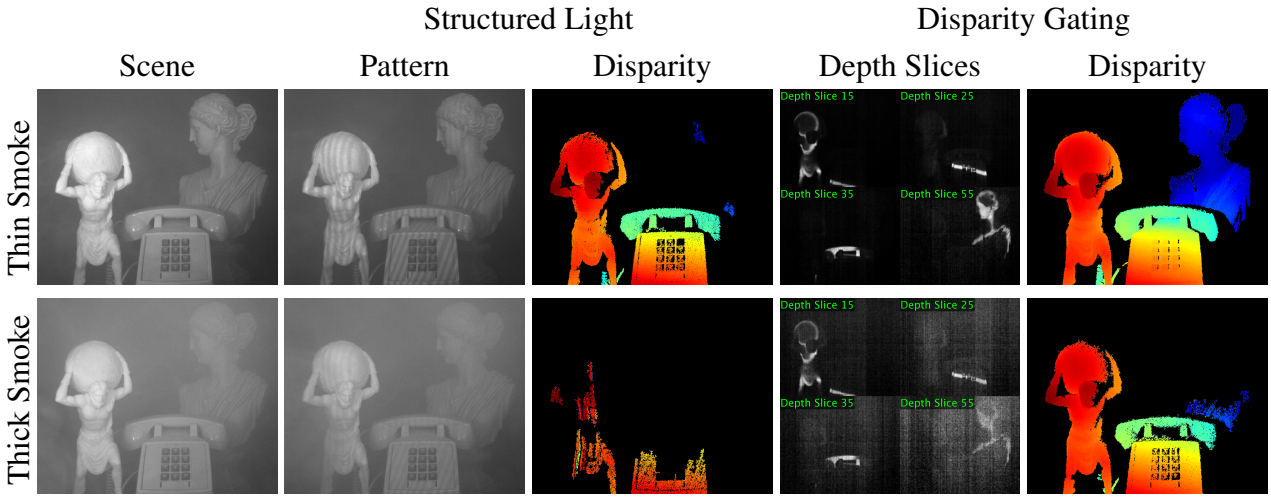


Figure 4.11: Disparity gating in participating media. We compare structured light with regular imaging to disparity gating. Disparity Gating is able to reconstruct objects at a further distance through the smoke than regular structured light. For structured light, we use 20 patterns and average 3 16.67ms exposures for each pattern. For disparity gating we divide the range of disparities in the scene into 60 slices and capture each slice with a 16.67ms exposure.

sensor's working principle. The performance of our sensor depends on the projector and cameras remaining aligned in a rectified configuration. We briefly study the sensitivity of the sensor to small alignment errors.

4.5.1 Range and Power

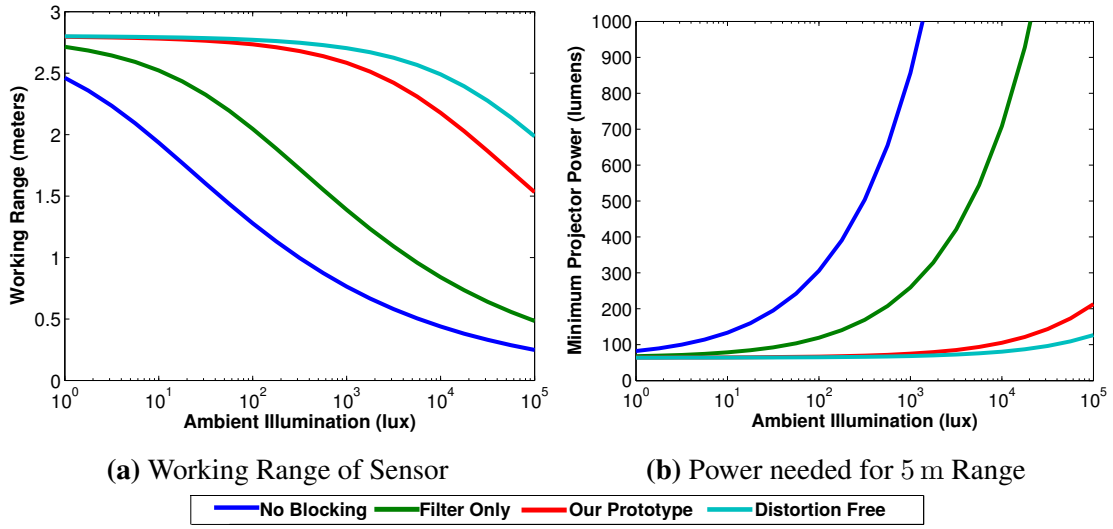
The working range of a active illumination system is limited by read noise and shot noise. The contribution of ambient light to the image is essentially independent of distance, while light from the active illumination source has an inverse square fall off. When the ambient light reaching the camera is small compared to the light from the active source, range is limited by read noise. As the relative contribution of ambient light to the image grows, the light from the source is lost in the shot noise of the ambient component.

Weakening the effect of ambient light by coding or optical filtering can increase the working range of an active illumination system at a given power level. Let k be the factor by which an imaging system can effectively weaken the influence of ambient light. Consider these four cases:

1. using no coding or optical filtering ($k = 1$)
2. using a narrow band light source and a filter on the camera ($k = 15$)
3. an idealized (distortion free) system that images in epipolar-only mode using an exposure time of $20\,\mu\mathrm{s}$ and a filter ($k \approx 800 \times 15 = 12,000$);
4. a system like ours where the exposure time has been extended to be $100\,\mu\mathrm{s}$ to accommodate the effects of distortion ($k \approx 2,500$).

As shown in Figure 4.12, systems with a higher ability to block ambient light show significantly slower degradation in working range as ambient light levels increase (we define working range as the maximum distance at which a binary projected pattern can be reliably decoded). To attain a working range of $5\,\mathrm{m}$ with little or no blocking of ambient light (low k values) requires a very powerful light source, while systems like ours with a high k value are much more energy efficient.

With the high ambient light blocking capability of epipolar-only imaging, camera sensitivity becomes the main limiting factor in determining system range (the intensity of the returned light signal falls off with distance and drops below the camera noise threshold). By using more sensitive cameras for imaging ($2/3''$ CMOS sensors instead of the $1/1.8''$ CMOS sensors we currently use) we would be able to built a system with a range of $5\,\mathrm{m}$ using a light source with an output of only 30 lumens.

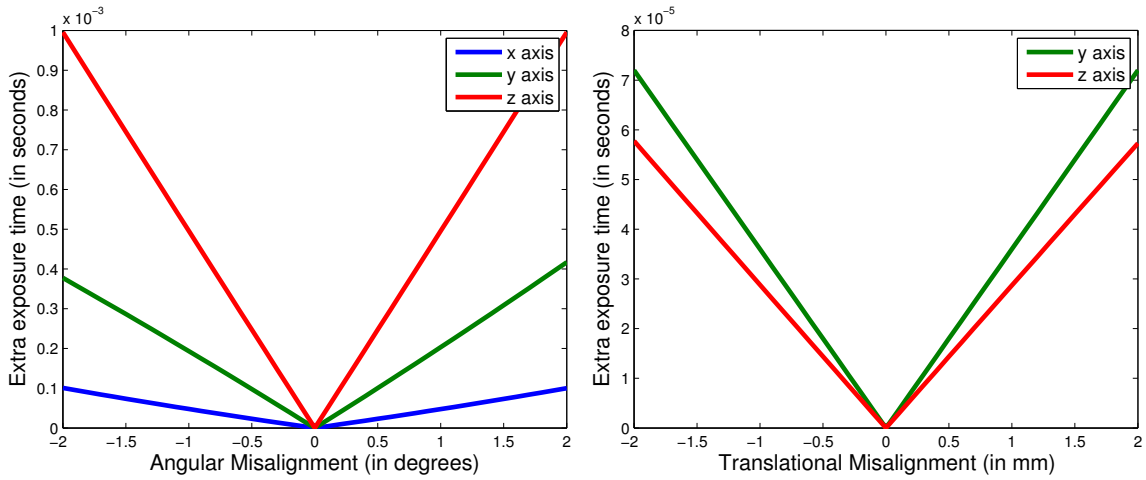


(a) Working Range of Sensor **(b)** Power needed for $5\,\mathrm{m}$ Range

Figure 4.12: Range analysis. (a) Maximum working range of a system with the same projector light output and cameras as our sensor as a function of ambient light level for different ambient blocking schemes. (b) The projector power required to attain a $5\,\mathrm{m}$ range. By effectively blocking a large fraction of the ambient light, epipolar-only imaging with a laser projector is able to provide energy-efficient operation even under bright ambient illumination conditions. Code listings for this analysis appear in the appendix.

4.5.2 Sensitivity to Alignment Errors

Like any triangulation-based system for depth measurement, the Episcan3D sensor needs a fixed, known baseline. With regular stereo systems if displacement between the cameras changes due to thermal deformations, shock, vibrations etc. the system can be realigned in software. With our sensor, the projector and camera need to be physically rectified because the masking mechanism we use to enable epipolar imaging requires each projector scanline to correspond to a row of camera pixels. Most alignment errors can only be partially corrected in software (the only controls over the camera are the trigger offset, exposure and clock speed). Figure 4.13 investigates how alignment errors would degrade performance. This analysis provides useful information about the level of alignment tolerance that is required when building the device.



(a) Rotational Errors **(b)**Translational Errors

Figure 4.13: Alignment Analysis: The Episcan3D sensor needs its projector and camera to be physically aligned in a stereo configuration. As the projector goes out of alignment with one of the cameras, the camera exposure needs to be increased for the epipolar light to be captured by the rolling shutter masks. When correctly aligned, our current prototype uses an exposure time of 100μs. Increasing the exposure time by another 100μs to provide slack for misalignment would double the amount of ambient light reaching the sensor. In (a) the slack exposure required for rotational misalignments along the various axes is plotted. The system is relatively less sensitive to rotations about the projector's x axis because these can be corrected for (to a point) in software by adjusting the trigger offset. In (b) we plot slack exposure required against translational misalignment. Misalignment along the x axis alters the baseline but does not affect rectification.

Chapter 5

Epipolar Time-of-Flight Imaging

"Don't guess. Measure. Measure twice."

5.1 Introduction

Depth cameras based on structured light and continuous wave time of flight (CW-ToF) are finding applications in a range of domains including 3D reconstruction [72], autonomous navigation, gesture recognition [93] and augmented reality. These depth cameras have high measurement rates and are inexpensive enough to be adopted widely in consumer applications. CW-ToF depth cameras are particularly attractive because they are compact (accuracy doesn't depend on a baseline between light source and sensor) and the processing needed to recover depth maps is minimal. However, compared to the more expensive impulse ToF sensors (scanning and flash LIDARs), CW-ToF depth cameras have poor robustness to ambient light, interference from other CW-ToF cameras and global illumination effects. Solving these problems would help make inexpensive depth cameras that can be deployed widely a reality.

The active light sources in CW-ToF depth cameras are limited by power and eye safety considerations. ToF sensors use pixel designs that electronically subtract out some or all of the DC component of the incident light [74] which helps prevent strong ambient sources (like sunlight) from saturating the sensor before enough light from the active source has been integrated. However, the more fundamental problem of photon shot noise induced by ambient light remains. The only way to reduce ambient shot noise is to concentrate the active source and shorten the exposure [13, 30, 65]. For a CW-ToF camera to operate accurately at long ranges in sunlight, some means of light concentration is necessary.

In addition to the direct one-bounce path via a scene point from light source to sensor, in most scenes there exists a continuum of multi-bounce, indirect paths (global illumination) between the source and sensor. As a result, each sensor pixel sees a mixture of path lengths which can result

in systematic errors in depth estimates. Significant errors can be caused by translucent surfaces, specular objects, surfaces illuminated at grazing angles and scattering media like fog.

In many applications, it is necessary for multiple depth cameras to be able to operate simultaneously in the same environment without having to synchronize or communicate with each other. CW-ToF cameras operating at the same modulation frequency corrupt each other's measurements. Using different modulation frequencies or specially designed coding schemes can prevent interference [9], but this requires cameras to use a common coding protocol and causes each depth camera to act as a source of ambient light (and hence noise) for the others. An additional means of preventing interference that is orthogonal to coding is desirable.

Although these three problems appear to be distinct, they all stem from a common cause - the fact that CW-ToF depth cameras typically illuminate and image the sensor's entire field of view at the same time. It is this parallelism that allows for their high measurement rates, but it also results in a lack of robustness. Spreading out light from the active source across entire field of view increases the likelihood that two CW-ToF cameras will interfere with each other, it increases the number of indirect light paths that contribute to the image, and it necessitates longer exposure times, increasing the effects of ambient shot noise.

Energy-efficient epipolar imaging which has been demonstrated in the context of non ToF, structured light systems [77] is an effective means of handling these three problems. Structured light epipolar imaging uses a scanning projector as the active light source. This source is placed in a rectified stereo configuration with the camera sensor. By the epipolar constraint, each row of projector pixels will correspond to one row of pixels on the sensor. The scene is illuminated and imaged one row at a time with the exposed row on the sensor synchronized to the active row of the scanning projector (see figure 5.3c).

With a scanning projector, light is concentrated into a small spatial extent at any point of time so a short exposure can be used for each row reducing the effects of ambient light. Because only a small subset of the possible paths between the light source and sensor are unblocked at any point of time, very few indirect light paths contribute to the captured image but all direct, one bounce light paths from source to sensor are captured because they obey epipolar geometry. Illuminating the scene one line at a time provides an effective compromise between the speed of full-field imaging and the robustness of point scanning. In microscopy, light sheet microscopy [106] offers a similar trade off between wide field and point scanning confocal techniques [67].

In this chapter, we extend epipolar imaging to continuous wave ToF depth camera systems. We analyze how epipolar imaging would improve the performance of a CW-ToF depth camera under bright ambient illumination, when multiple devices interfere and when multi-path interference occurs. We demonstrate the first live, energy-efficient implementation of epipolar CW-ToF imaging. This new epipolar imaging implementation provides the flexibility to chose the order

and frequency at which lines in the scene are captured.

Our epipolar ToF prototype is able to recover depth maps outdoors in bright sunlight ($70\,\mathrm{klx}$ at ranges of over $10\,\mathrm{m}$. Its ambient suppression abilities are so strong that it can obtain accurate depth returns from light bulbs even when they are turned on. It suffers from minimal interference from other CW-ToF devices, even if they are operating at the same modulation frequency. It is unaffected by many of the global illumination effects observed in everyday scenes and can recover live depth maps in these environments without placing any restrictive assumptions on the types of global light paths present in the scene.

5.2 Continuous Wave Time of Flight

The operating principles of CW-ToF cameras are discussed in [56]. To summarize, these cameras use a temporally-modulated light source and a sensor where the exposure is also modulated during integration. If the illumination modulation function is $f_\omega(t) = \cos(\omega t)$ and the sensor modulation function is $g_{\omega,\phi}(t) = \cos(\omega t + \phi)$ where ω is the modulation frequency in rad/s and ϕ is the phase offset between the source and sensor modulation functions, then the measurement integrated over an exposure time of t_{exp} at a pixel x is

$$I_{\omega,\phi}(x) = \int_0^{t_{\text{exp}}} f_\omega(t) * [h_x(t) + A_x] g_{\omega,\phi}(t) dt \tag{5.1}$$

$$= \tfrac{t_{\text{exp}}}{2} \int_0^{\infty} \cos\left(\omega\tau - \phi\right) h_x\left(\tau\right) d\tau \;, \tag{5.2}$$

where $*$ denotes convolution, $h_x(t)$ represents a pixel's transient response to the active light source and A_x is the response due to the DC component of the active light source as well as other ambient sources. In practice, $I_{\omega,\phi}(x)$ is measured by integrating incoming light to two different storage sites (called taps) depending on whether $g_{\omega,\phi}(t)$ is positive or negative and then taking the difference between the stored values. Thus even though A_x drops out of the integral, ambient light still adds to the measurement shot noise.

If there are no indirect light paths between the light source and sensor pixel x, then $h_x(t) \propto \delta\left(t - {}^{l\,(x)}/_{c}\right)$ where c is the speed of light and $l(x)$ is the length of the path from the light source to the scene point corresponding to x and back to the sensor. Assuming the scene is static, we can recover the path length $l(x)$ by capturing a pair of images at the same frequency but two different modulation phases $\phi = 0$ and $\phi = \pi/2$:

$$l(x) = \frac{c}{2\omega} \operatorname{atan2}\left(I_{\omega,\frac{\pi}{2}}(x), I_{\omega,0}(x)\right) \;. \tag{5.3}$$

The pixel depth $z(x)$ can be computed from $l(x)$ using the geometric calibration parameters of the light source and sensor.

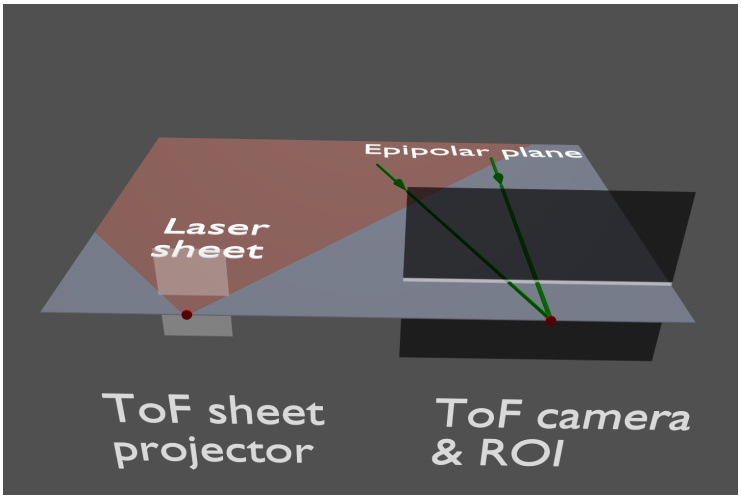
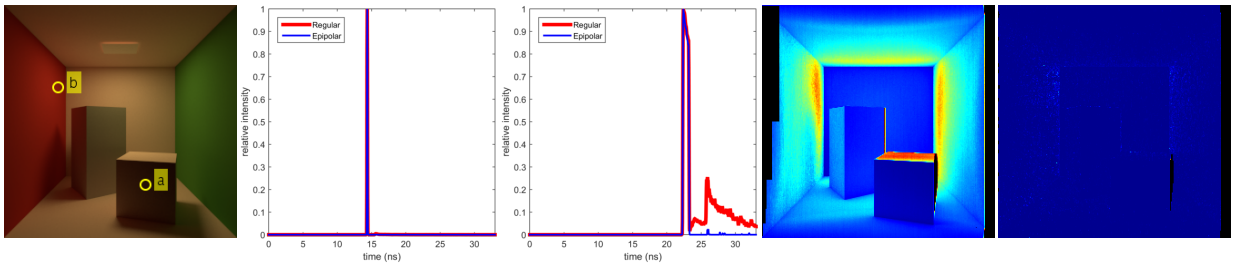


Figure 5.1: Epipolar time of flight. A projector that generates a steerable sheet of modulated laser light is combined with a ToF sensor whose rows can be exposed one at a time. The projector and camera are placed in a rectified stereo configuration so that the light sheet always lies on an epipolar plane between the projector and the camera. At any given instant, only the row of camera pixels on the epipolar plane is exposed to light..

In scenes with global illumination, the transient function $h_x(t)$ contains components in addition to the Dirac delta of the direct light path and applying Eq. (5.3) causes depth estimate errors. A common solution is to use different modulation frequencies and fit a more complex model of $h_x(t)$ to the resulting measurements. In [18, 47] a series of Dirac deltas are fitted to $h_x(t)$ to model specular inter reflections and reflections from translucent surfaces. Mixtures of exponentially modified Gaussians have been used to model volumetric [40] and subsurface [97] scattering. The disadvantage of these approaches is that the chosen model needs to be a good fit to the multi-path effects present in the scene. The requirement for more measurements increases acquisition time.

ToF illumination and imaging codes can be designed to minimize certain types of global illumination effects. By using very high modulation frequencies, diffuse interreflections can be removed [31]. The depth selective camera [100] uses codes that can capture light corresponding to particular path lengths. In the context of handling global illumination, the method closest to our proposed epipolar CW-ToF scheme is [76] where the epipolar ToF image is computed by taking the difference of a regular ToF image and a non-epipolar ToF image. However, their implementation suffered from poor light efficiency and required thousands of frame captures to synthesize the non-epipolar ToF image which precluded its use in a live context. Also, it required long exposure times and was dependent on digital image subtraction so would have been unable to handle bright ambient light or interference from other devices.



(a) Scene (b) Transient at a (c) Transient at b (d) Regular Error (e) Epipolar Error

Figure 5.2: Simulation showing the effects of of global illumination on continuous wave time-of-flight with regular and epipolar imaging. (a) a Cornell box scene - for scale, the back wall is roughly $3.7\,\mathrm{m}$ from the camera. (b) at point a - on the cube in the foreground, there is little global illumination so the transient responses under both regular and epipolar imaging are close to impulse responses. (c) at point b - near the corner on the red wall, there is a strong inter-reflection which shows up as a long tail in the transient response under regular imaging which is largely suppressed in epipolar mode. (d) shows the difference between the groundtruth depth and depth estimated using regular imaging CW ToF (modulation frequency $30\,\mathrm{MHz}$) as a heatmap (scaled between $0\,\mathrm{cm}$ and $50\,\mathrm{cm}$), there are large errors due to interreflections. (e) the same error visualization for epipolar imaging, almost all the errors are suppressed.

5.3 Implementing Epipolar ToF

An epipolar ToF depth camera that is optimal in terms of light efficiency and usage of sensor bandwidth requires a light source that can concentrate its power into a single plane and a ToF sensor that can implement exposure coding to allow for only a single row of pixels to be exposed at a time. These two components must be placed in a rectified stereo configuration as shown in figure 5.1.

To realize the geometry of Figure 5.1, we use a line laser source with a 1D-scanning mirror that projects a steerable light sheet onto the scene. No currently-available CW-ToF sensor provides controllable exposure coding across the 2D pixel array. Taking into account available off-the-shelf hardware, there are three possible ways to restrict exposure to pixels on a single epipolar plane:

1. Use a Digital Micromirror Device (DMD) to mask all other pixels
2. Use a 1D sensor and a controllable mirror to select the epipolar plane it should image
3. use a 2D sensor with a controllable region of interest (ROI).

We chose the third option because it is much more light-efficient than a DMD mask and leads to a much simpler design. We make the ROI one row tall to match the requirements of epipolar

ToF.

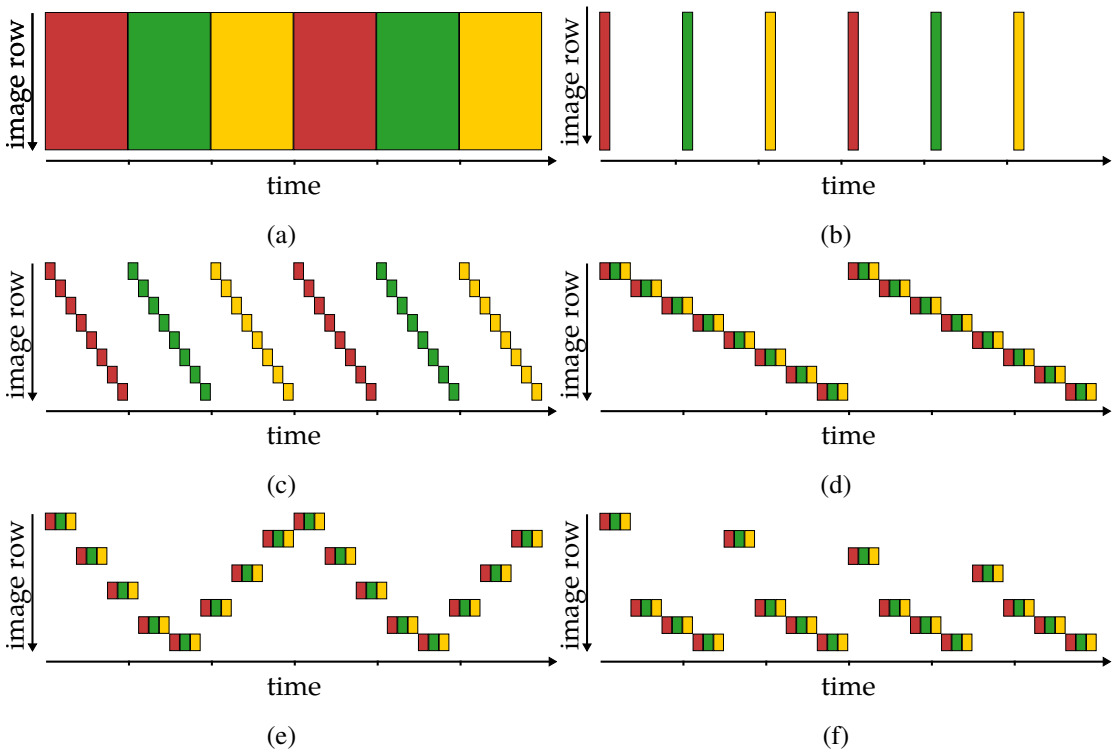


Figure 5.3: Epipolar plane sampling schemes and row exposures in ToF imaging. (a) In conventional CW-ToF all epipolar planes are illuminated simultaneously and all camera rows are exposed at the same time. This requires long exposures and leads to severe artifacts due to motion, ambient light, global light transport and interference between devices. (b) Sending a very brief, high-intensity pulse of light for CW-ToF confers resistance to ambient light but is still prone to artifacts due to global light transport and motion. (c) Ordering the epipolar ToF planes similarly to a rolling-shutter camera confers robustness to ambient light, global illumination and motion blur. Sensitivity to motion remains, however, because of the significant delay between the phase measurements acquired for each row. (d) Interleaving those measurements plane by plane minimizes such artifacts. (e) Scanning the entire field of view twice within the same total exposure time yields higher temporal sampling of the scene and makes consistent merging of individual depth map rows easier. (f) For certain applications, scanning different portions of the field of view with different temporal sampling rates can be beneficial.

5.3.1 Epipolar plane sampling

As explained in Section 5.2, CW-ToF requires at least two images to recover depth. To cover an entire scene using epipolar ToF, the active epipolar plane must be swept across the field of view. This offers flexibility to chose the order in which epipolar planes are sampled.

Figure 5.3 illustrates several such ordering schemes. For instance, the ordering scheme of Figure 5.3c illustrates the operation of a hypothetical rolling-shutter ToF camera, where one complete image is acquired for each modulation phase. This scheme is undesirable because if the scene or camera move while acquiring these images, the recovered depth map will contain hard-to-correct errors.

A better ordering strategy is to loop through the set of modulation phases at one epipolar plane before imaging the next row (Figure 5.3d). Since each row's exposure time is very short, all phases required for a single row can be acquired quickly enough to minimize depth and motion blur artifacts due to camera/scene motion.

Under this strategy, each row is captured at a slightly different time. Although this induces a rolling-shutter-like effect in the acquired depth map, the individual depth values will be blur- and artifact-free and can be combined into a consistent model by post-processing [4, 49].

To make such post-processing even easier while obeying the kinematic constraints of the mirror's actuator, we order epipolar planes in a sawtooth pattern (Figure 5.3e). This essentially provides full-field-of-view depth maps at twice the frame rate but half the vertical resolution, making depth correction easier for fast camera shake and/or scene motions. More generally, Figure 5.3f shows an example of a non-uniform sampling scheme in which epipolar planes corresponding to lower image rows are sampled more frequently. This type of sampling could be useful on a vehicle where lower portions of the field of view are usually closer and move faster, requiring acquisition at a faster sampling rate.

5.4 Hardware Prototype

Our prototype device for epipolar time-of-flight imaging uses a galvomirror-based light sheet projector for illumination and a ToF sensor with an adjustable region of interest for imaging.

The time-of-flight sensor we use is the EPC660 (from Espros Photonics) which has a resolution of 320x240 and pixels that implement ambient saturation prevention. The sensor is fitted with a $8\,\mathrm{mm}$ F1.6 low distortion lens and an optical bandpass filter ($650\,\mathrm{nm}$ center frequency, $20\,\mathrm{nm}$ bandwidth). The sensor allows the ROI to be changed with every sensor readout and we use this feature to select which row to image. We read data out of the sensor using the sensor development kit (DME660) from the manufacturer.

Our line projector uses a $638\,\mathrm{nm}$ laser diode with a peak power of $700\,\mathrm{mW}$. Light from the

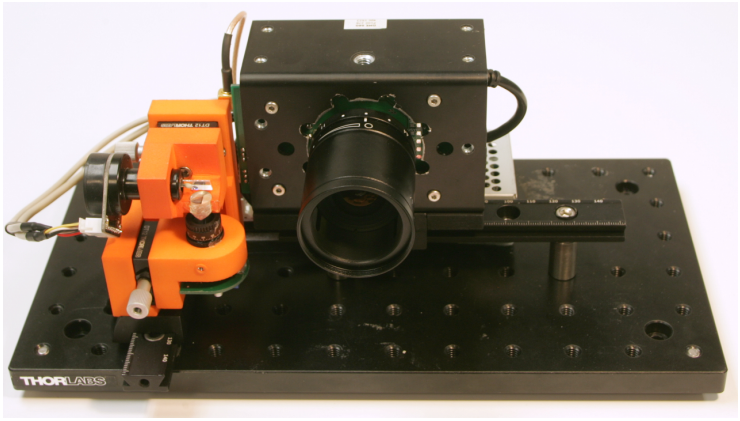


Figure 5.4: Our Epipolar ToF prototype uses a custom built steerable light sheet projector (orange assembly) and a DME660 camera with fast ROI control to capture arbitrary rows of pixels.

diode is collimated and passed through a Powell lens that stretches the beam cross-section into a diverging, almost uniformly illuminated straight line with a $45°$ fanout angle. The laser light is directed at a 1D scanning galvomirror that can be rotated to deflect the sheet. The rotational range of the mirror gives the projector a $40°$ vertical field of view. The projector's effective center of projection moves as the mirror rotates but this effect can be ignored because the distance between the fanout point and the galvomirror is very small compared to depths in the scene.

A microcontroller is used to synchronize the sensor and light source. The microcontroller communicates with the sensor over an I2C bus to set the exposure time, modulation frequency/phase and region-of-interest row and also to trigger each capture. The microcontroller also actuates the projector's galvomirror. In addition, the microcontroller can read the camera's rotational velocity from a MEMS IMU (inertial magnetic unit) that we have attached to the sensor. A frequency generator circuit allows us to select a modulation frequency between $1\,\mathrm{MHz}$ and $24\,\mathrm{MHz}$ in steps of $1\,\mathrm{MHz}$.

We align the projector and camera side-by-side in a rectified stereo configuration as required for epipolar imaging. When correctly aligned, the projected light sheet illuminates a single row of pixels in the camera and this row is independent of depth. A mirror calibration is performed to determine the mapping between galvomirror angle and illuminated camera row.

5.4.1 Sensor Calibration

In practice, we observe that the measurements read out from the sensor do not match their expected values. There are a number of reasons for this discrepancy, including fixed-pattern noise, non-uniform pixel sensitivity, crosstalk between taps and small variations in the phase of the

exposure modulation function at each pixel. We model the relation between the expected sensor measurements $I_\omega(x)$ and the actual measurements $\hat{I}_\omega(x)$ using a 3×3 correction matrix $H_\omega(x)$ at each pixel

$$\begin{bmatrix} I_{\omega,0} \\ I_{\omega,\frac{\pi}{2}} \\ 1 \end{bmatrix} = H_\omega(x) \begin{bmatrix} \hat{I}_{\omega,0} \\ \hat{I}_{\omega,\frac{\pi}{2}} \\ 1 \end{bmatrix} \tag{5.4}$$

To find $H_\omega(x)$, we place a fronto-parallel surface at a set of known distances z_k, $k = 1, .., K$. For each position of the plane, we collect sensor measurements at different aperture settings ($s = 1, ..., S$) to simulate the effect of varying scene albedos. For each plane position k, we can compute the path length at a pixel $l_k(x)$ and thus the expected phase $\frac{2\omega l_k(x)}{c}$. We then compute the correction matrix that best explains the sensor measurements $I_{\omega,k,s}(x)$ by finding the matrix that minimizes the least-squares error between the corrected measurements and the expected phase.

These calibration parameters are dependent on both modulation frequency and exposure time so we repeat the process for all the frequencies and exposure times used in the experiments. Although the modulation signals we input to the sensor and light source driver are square waves, at modulation frequencies of $20\,\mathrm{MHz}$ and above, we noticed that the harmonics were largely suppressed and so the modulation functions were well approximated by sinusoids.

5.4.2 Timing

The time needed to image a row (and by extension the frame rate) with our prototype is a function of several quantities: the number of readouts per row n, the exposure time t_{exp}, the readout time for a row t_{read} and the time t_{mirror} taken by the galvo mirror to move to the next row position in the sampling sequence:

$$t_{\text{row}} = n t_{\text{exp}} + (n-1) t_{\text{read}} + \max(t_{\text{read}}, t_{\text{mirror}}) \tag{5.5}$$

With a two-tap sensor like the one in our prototype, at least $n = 2$ readouts are needed to measure depth using a single modulation frequency. Figure 5.5 shows a timing example. t_{read} is $175\,\mu\mathrm{s}$ and for most of our experiments we set t_{exp} to $100\,\mu\mathrm{s}$. In our row sampling sequence, the mirror rotates through two rows (approximately $0.33°$) per step and t_{mirror} for this step size is roughly $100\,\mu\mathrm{s}$. In total, t_{row} is $550\,\mu\mathrm{s}$ when $n = 2$ yielding a frame rate of 7.5fps (or 3.8fps when $n = 4$).

5.4.3 Limitations

Currently, the main bottleneck for frame rate is readout time. Our approach needs data from only one row of the sensor per readout, but the smallest region of interest the EPC660 sensor supports

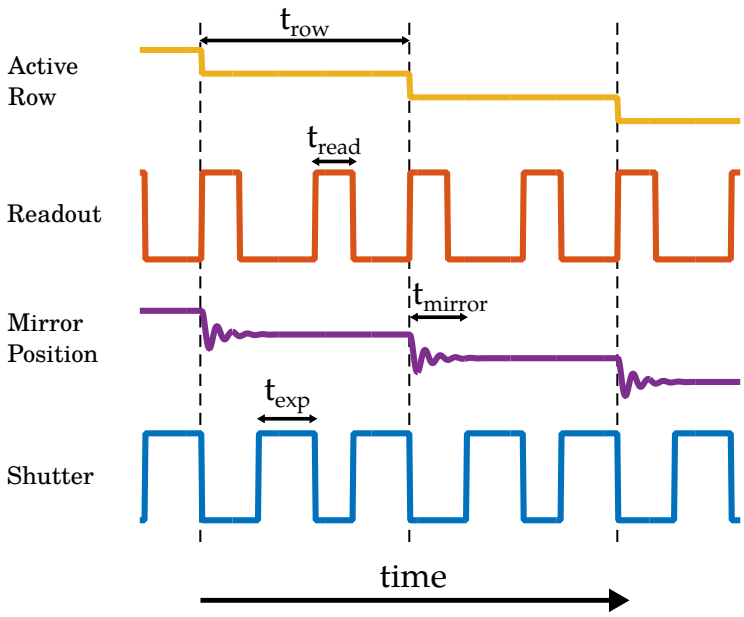


Figure 5.5: Timing diagrams for camera exposure, readout and mirror position for a particular sequencing of the rows. First, the scanning mirror is moved to the new active row and takes t_{mirror} to settle in the position. When the previous row readout is complete (which takes t_{read}) and the the mirror is in position, the camera is triggered. Each exposure lasts for t_{exp} and at the end of each exposure the row is read out.

is 4 rows tall. We are forced to read out 4 rows at a time when we actually use just one. Also, the development kit we have used limits the sensor data readout bus to $20\,\mathrm{MHz}$ but the sensor itself supports bus rates up to $80\,\mathrm{MHz}$. The minimum value of t_{exp} depends on the light source's peak power and desired range. Our prototype has a source with a peak power of $700\,\mathrm{mW}$ while most other experimental time-of-flight systems have a peak light source power in the $3\,\mathrm{W}$ to $10\,\mathrm{W}$ range. With a brighter light source we could use a far shorter exposure time without loss of range. Lastly, the low-cost galvomirror we used could be replaced with a faster 1D MEMS mirror. With these improvements a system based on our prototype would operate at video frame rates.

The sensor used in our prototype supports a maximum modulation frequency of only $24\,\mathrm{MHz}$ whereas most other time-of-flight sensors can run in the $50\,\mathrm{MHz}$ to $100\,\mathrm{MHz}$ range. This limits our prototype's ability to accurately scan smaller objects or be used for transient imaging. The EPC660 datasheet specifies that the sensor ADC returns 12-bit values but the version of the sensor we are using only returns 10 bits. This affects range and makes the output depth maps noisier.

5.4.4 Eye Safety

Eye safety requirements place a limit on the power that can be emitted by a CW-ToF system's light source. This has implications for accuracy, range and frame rate. The quantity of interest in determining eye safety for a laser source is the Maximal Permissible Exposure or MPE. MPE is expressed in terms of energy or power per unit area [5] and is function of light source wavelength and exposure time among other factors. In our laser sheet projector, light spreads out from a spot so the power density drops as the distance from the source increases. For our current system, the energy density is safe at a distance of at least $66\,\mathrm{cm}$ from the source. By switching to a near-infrared ($850\,\mathrm{nm}$) laser, the eye safe distance of our system can be reduced to $40\,\mathrm{cm}$. Details of the calculation are listed in Table 5.1. The laser diode source we currently use is effectively a point source. The permissible energy limits for extended light sources are considerably higher. Switching to a small extended area source such as a Vertical-Cavity Surface Emitting (VCSEL) array, would allow us to make our prototype eye safe at shorter distances and/or extend the maximum working range.

5.5 Results

We demonstrate the benefits of epipolar ToF imaging by comparing to regular ToF imaging in different scenes and conditions. There are two ways we could implement regular ToF imaging

Quantity	Symbol	Formula	Value (Visible Light)	Value (Near-Infrared)
laser peak power	P		$700\,\mathrm{mW}$	$700\,\mathrm{mW}$
laser wavelength	λ		$638\,\mathrm{nm}$	$638\,\mathrm{nm}$
laser spot diameter	d		$0.4\,\mathrm{cm}$	$0.4\,\mathrm{cm}$
laser horizontal fanout	θ_x		$0.785\,\mathrm{rad}$	$0.785\,\mathrm{rad}$
laser vertical fanout	θ_y		$0.002\,\mathrm{rad}$	$0.002\,\mathrm{rad}$
eye-laser distance	z			
illuminated width	$w(z)$	$d + z\theta_x$		
illuminated height	$h(z)$	$d + z\theta_y$		
illuminated area	$A(z)$	$w(z)h(z)$		
exposure duration	t_{exp}		$100\,\mu\mathrm{s}$	$100\,\mu\mathrm{s}$
energy per exposure	E	$1/\sqrt{2}Pt_{\text{exp}}$	$49.5\,\mu\mathrm{W}$	$49.5\,\mu\mathrm{W}$
energy density limit	MPE		$1.80\,\mu\mathrm{J/cm^2}$	$3.59\,\mu\mathrm{J/cm^2}$
eye safe distance	z_s	$A(z_s) = \frac{E}{MPE}$	$66\,\mathrm{cm}$	$40\,\mathrm{cm}$

Table 5.1: The maximal permissible exposure (MPE) is the highest safe energy density for a light source. The rules for computing MPE depend on exposure time and wavelength amongst other factors. Our current prototype is eye safe at distances greater than $66\,\mathrm{cm}$. A near-infrared ($850\,\mathrm{nm}$) version of our current prototype would be eye safe at distances greater than $40\,\mathrm{cm}$. We designed the laser sheet projector so that its horizontal fanout angle matches the camera's horizontal field of view and the vertical fanout angle is slightly less than the angle covered by a camera pixel.

Scene	Albedo	0.5
	Depth	$10\,\mathrm{m}$
Projector	Peak Power	$2\,\mathrm{W}$
	Output wavelength	$850\,\mathrm{nm}$
	Modulation frequency	$15\,\mathrm{MHz}$
Camera	Focal length	$8\,\mathrm{mm}$
	Resolution	320×240
	f-number	1.1
	pixel size	$20\,\mu\mathrm{m} \times 20\,\mu\mathrm{m}$
	filter bandwidth	$20\,\mathrm{nm}$
	Quantum efficiency	0.8
	read noise	$5e^-$
	electrons per LSB	1

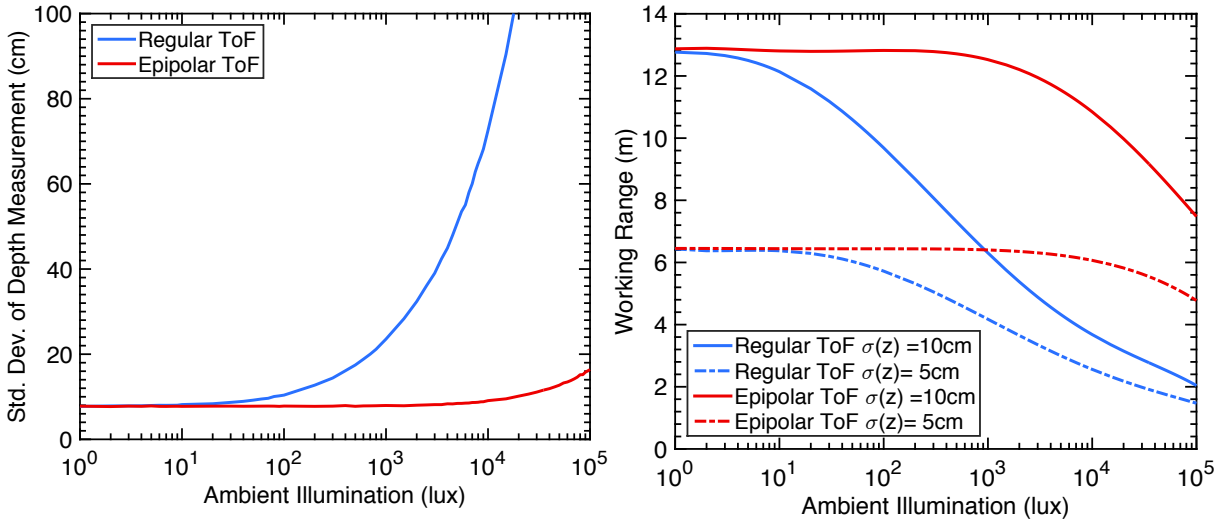
Table 5.2: Parameters for Simulated ToF Camera

with our prototype sensor. The first is to remove the galvomirror and the line generator lens from the laser sheet projector and replace them with a diffuser. The second is to keep the entire sensor exposed until the sheet projector has swept across the full field of view. In the multi-device interference and camera motion experiments, we use a diffuser. For the ambient light comparisons, we use the full frame ROI approach. This prevents light loss at the diffuser from affecting our comparisons.

5.5.1 Ambient Light

Figure 5.6 shows a simulation that illustrates the benefits of applying epipolar imaging to ToF in brightly lit environments. For a given light source power, depth accuracy degrades rapidly with regular imaging as ambient light levels increase from $0\,\mathrm{lx}$ (complete darkness) to $100\,\mathrm{klx}$ (direct sunlight). With epipolar imaging, the degradation is much more gradual. Figure 5.6 simulates an almost idealized camera where the only noise effects are photon shot noise and a very small amount of read noise. Dark noise and ADC quantization effects are not considered. See table 5.2 for a list of parameters used.

Figure 5.7 quantitatively compares our sensor prototype operating outdoors in regular ToF and epipolar ToF imaging modes under cloudy and sunny conditions. Regular ToF mode performs poorly in bright sunlight, while epipolar ToF is considerably more robust. Figure 5.9 shows an example scene with both strong ambient light and global illumination effects.

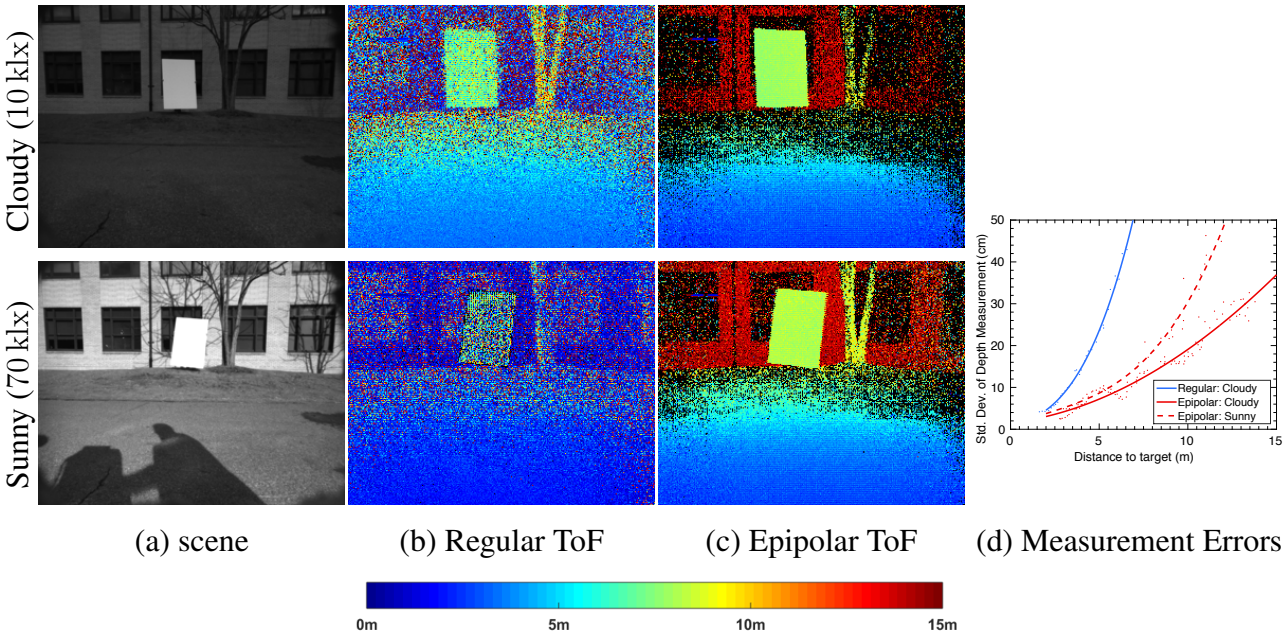


(a) (b)

Figure 5.6: A simulation of the standard deviation in depth measurements obtained using regular and epipolar ToF imaging ($15\,\mathrm{MHz}$ modulation frequency) for a target $10\,\mathrm{m}$ from the camera as a function of ambient light level is shown in (a). For both cases, the peak light source power is $2\,\mathrm{W}$ and the total exposure time is the same ($7.2\,\mathrm{ms}$ per image) but epipolar ToF is more robust to ambient light because it concentrates light source power and uses a short exposure for each row ($30\,\mu\mathrm{s}$). (b) shows the working range of the same simulated camera at different levels of acceptable range accuracy. Note that simulated camera's parameters differ from prototype, see table 5.2 for parameters and appendix for code listings.

5.5.2 Global Illumination

Figure 5.10 demonstrates the ability of epipolar imaging to suppress the effects of global illumination in a few common indoor environments. These results are generated using a single modulation frequency ($24\,\mathrm{MHz}$). In regular ToF mode, diffuse interreflections between the walls and ceiling cause depths to be overestimated and the corner to be rounded out. With epipolar imaging, the walls appear straight and meet at a sharp right angle. The conference table in the second row appears specular at grazing angles. In the bathroom scene, the ghosting on the wall due to reflections from the mirror is suppressed by epipolar imaging. The water fountain is particularly challenging because the direct return from its metallic surface is very weak, but the surface reflects a lot of indirect light back to the sensor. For epipolar imaging, we combine 3 exposures to try recover a usable direct signal. Longer exposures do not help regular imaging because the interreflections cause the sensor to saturate.



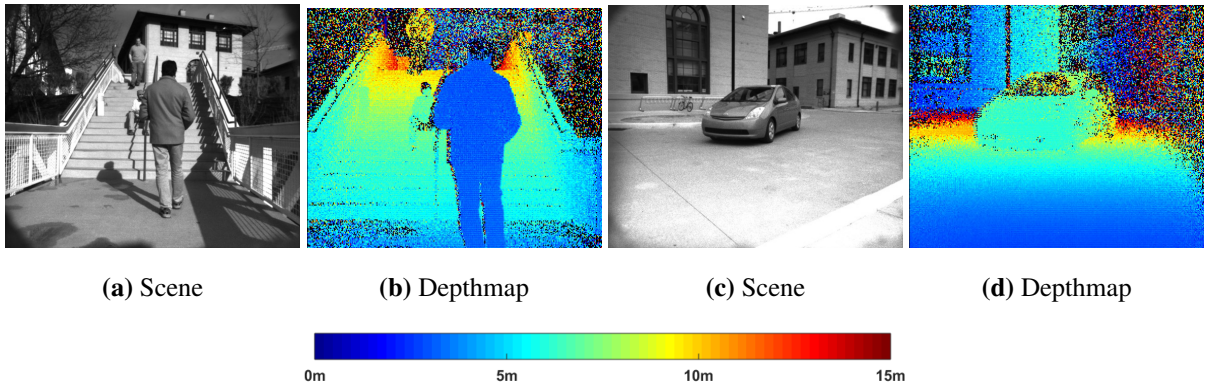
(a) scene (b) Regular ToF (c) Epipolar ToF (d) Measurement Errors

Figure 5.7: We placed a white planar target at a range of distances from the sensor in both cloudy weather and bright sunshine. Even under cloudy conditions, epipolar ToF imaging produced far less noisy depth measurements than regular ToF. Under bright sunlight, regular ToF failed completely whereas epipolar ToF still provided useful depth returns. The camera modulation frequency was set to $10\,\mathrm{MHz}$. (d) shows standard deviation in depth estimates versus distance to target (slower rising curves are better). Our prototype has depth error of around 3% at $10\,\mathrm{m}$ in bright sunlight.

5.5.3 Multi-Camera Interference

With epipolar CW-ToF imaging, two systems running at the same modulation frequency can usually only interfere with each other at a sparse set of pixels in each image. Each system illuminates and images a single line in the scene at a time, so at any instant the second system can only interfere with the first at the points where its illuminated line intersects with the first system's exposed row of pixels. A degenerate case occurs when two systems happen to be aligned in such a way that they have identical epipolar planes and their cameras are synchronized by chance. This, however be considered a very rare occurrence.

If more than two epipolar ToF systems are present, each pair of cameras has only a sparse set of pixels that may be affected by interference. When a set of epipolar ToF systems are running at different modulation frequencies, the contribution of each system to shot noise in the others is greatly reduced. Figure 5.11 shows the result of operating two CW-ToF cameras simultaneously at the same frequency in either regular or epipolar imaging modes. In epipolar mode, the inter-



(a) Scene **(b)** Depthmap **(c)** Scene **(d)** Depthmap

Figure 5.8: Depthmaps from our epipolar ToF camera outdoors on a sunny day. The camera modulation frequency is $10\,\mathrm{MHz}$ and so the maximum ambiguity-free range is $15\,\mathrm{m}$. The resulting wrap around in depth values is noticeable in the wall behind the car in (d) which is more than $15\,\mathrm{m}$ away.

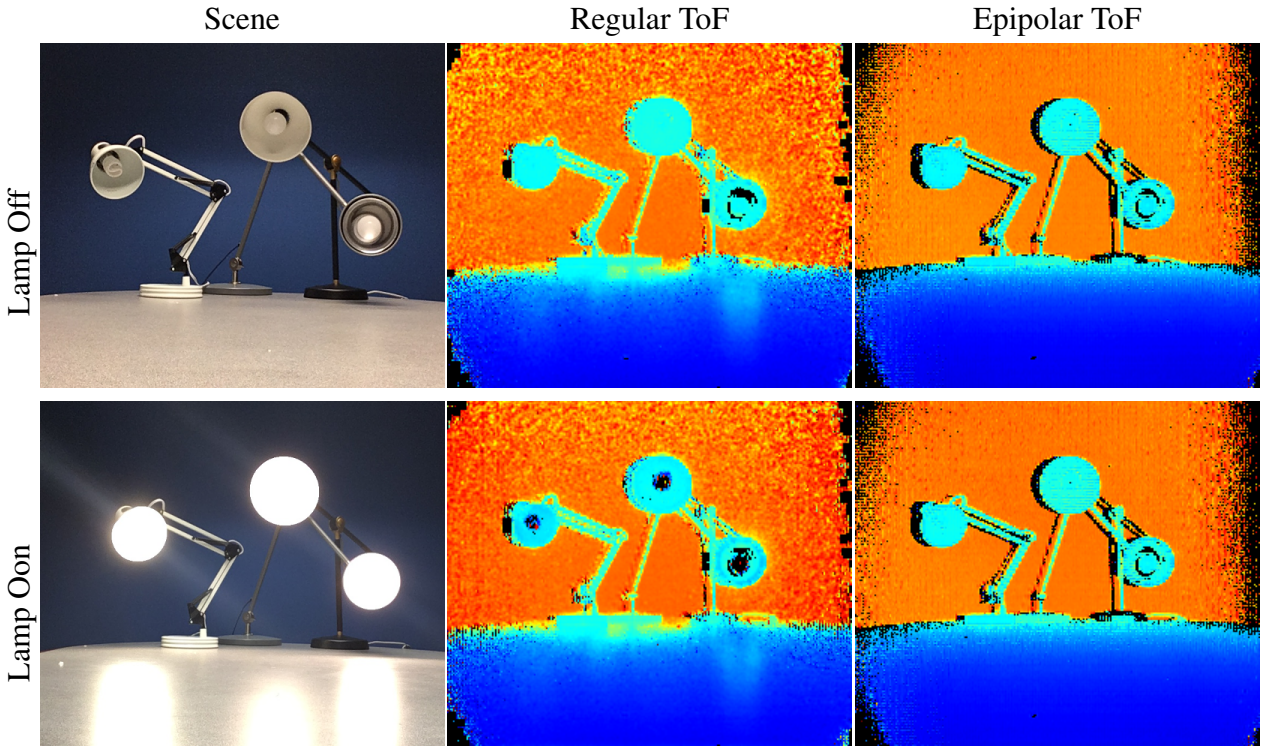
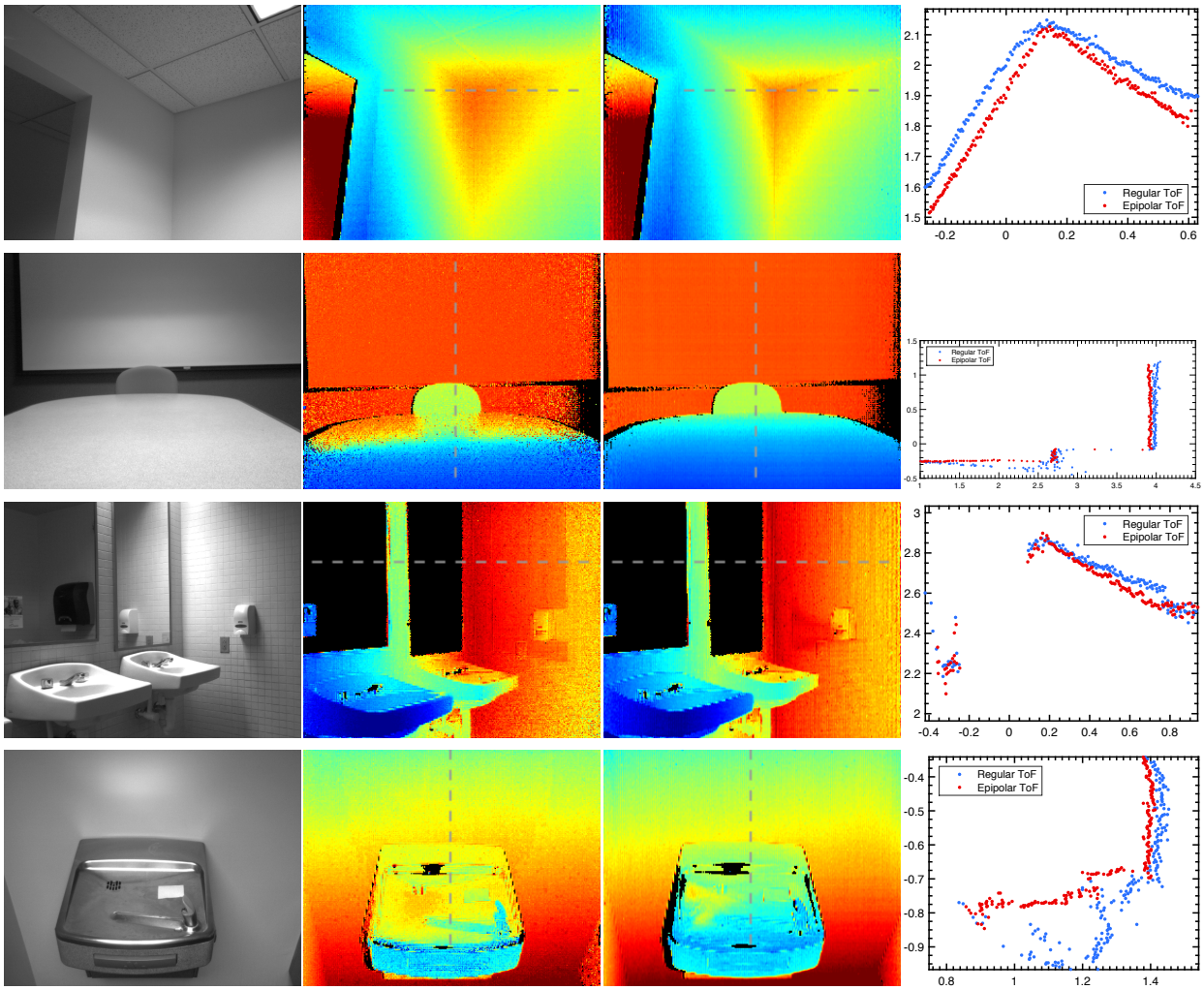


Figure 5.9: Epipolar ToF imaging provides accurate depth measurements from the surface of the light bulbs even when they are turned on. Also note how reflections from the table's surface cause errors with regular ToF, but these are suppressed with epipolar imaging.



(a) scene (b) Regular ToF (c) Epipolar ToF (d) Profile

Figure 5.10: Comparing depth maps with epipolar and regular ToF imaging in the presence of global light transport: diffuse interreflections at the corner, glossy interreflection from projection screen onto a shiny conference table, reflections from the mirrors in the restroom and in between the wall and the shiny water fountain. Epipolar ToF eliminates most of the global light transport resulting in depth maps that are significantly more accurate than regular ToF. All profile curves are in meters.

ference between the cameras is minimal. It should be noted that the two cameras are operating completely independently of each other without any form of synchronization between them.

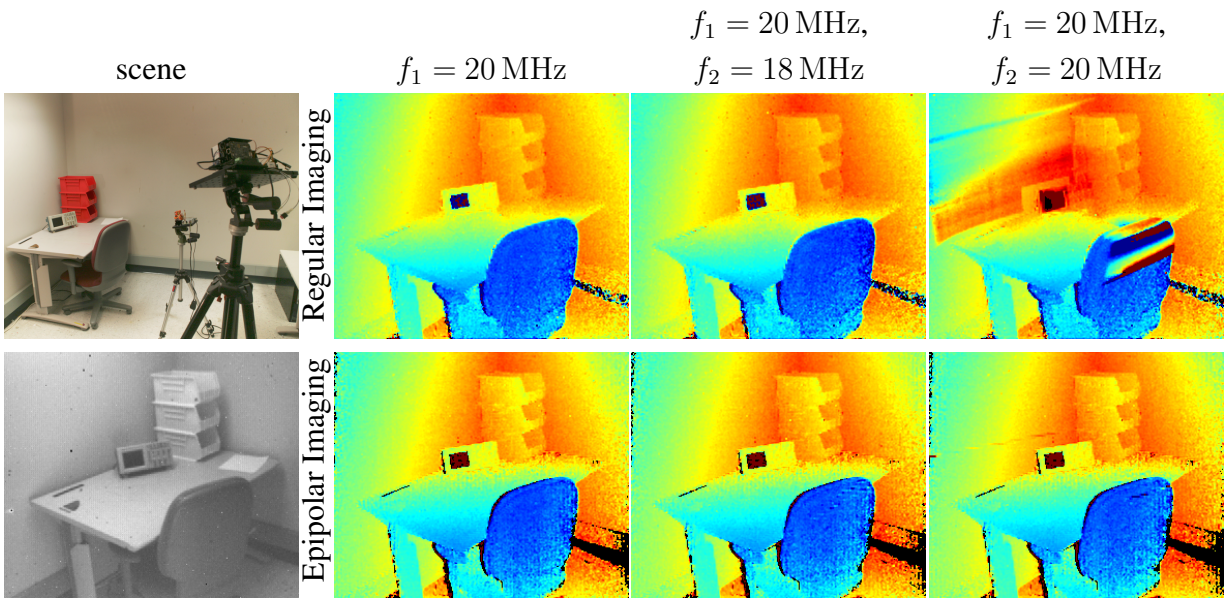


Figure 5.11: Two time-of-flight cameras are placed in a scene. When only one camera is running (second column), both regular imaging and epipolar imaging produce accurate depth maps. With both cameras operate at different frequencies (third column), the depth map with regular imaging becomes noisier because the other camera's light source adds to the shot noise while epipolar imaging is almost unaffected. When both cameras operate at the same modulation frequency (last column), they interfere strongly in regular imaging mode. With epipolar imaging, the two cameras will usually either not interfere at all, or will interfere along a single line through the image.

5.5.4 Camera Motion

Consider the case of a rotating camera with known rotational trajectory obtained from a MEMS gyroscope. With regular imaging, each captured ToF measurement has motion blur and strong artefacts at depth discontinuities because the measurements are not aligned to each other. This could be partially corrected using a spatially varying deconvolution but high frequencies in the image would be recovered poorly. With epipolar ToF imaging, motion blur has basically no effect and a depth map with a rolling-shutter-like effect is acquired. This can be corrected with a simple image warp computed from the rotation. Figure 5.12 shows an example from a rapidly panning camera.

(a) scene (b) Regular ToF (c) Epipolar ToF (d) Epipolar ToF (corrected)

Figure 5.12: With regular ToF, fast camera motion causes blur and misalignment between successive images. This causes errors in the recovered depth map. Even with known rotation, this is hard to fix because it involves deconvolving the captured images to register them (the convolution kernels are known, but the deconvolution process smooths over a lot of detail). With epipolar ToF (c), motion blur is minimized and there are no depth errors, but an interleaved rolling shutter like artefact is introduced by our coded exposure scheme. This can be corrected using a simple image warp on the depth map to give a corrected result (d)

5.6 Discussion

Epipolar imaging for continuous-wave time-of-flight depth cameras mitigates many of the problems commonly encountered with these sensors. These problems include highly degraded performance in brightly lit conditions, systematic errors due to global illumination, errors due to inter-device interference and artifacts induced by sensor motion.

Cycling through multiple phases or patterns at one row before proceeding to the next row is directly applicable to structured light as well. Such a scheme would make it possible to apply multi-image structured light methods to dynamic scenes for generating high-quality depth maps where currently only single-shot methods can be used.

In our prototype, the scanning mirror follows a sawtooth pattern and captures rows in an ordered sequence. However, with a faster scanning mirror, pseudo-random row sampling strategies could be implemented that might allow epipolar imaging to be used in conjunction with compressed sensing or similar techniques. This would allow recovery of temporally super-resolved depth maps of fast-moving scenes.

Chapter 6

Analysis

An important taxonomy of active illumination devices is on the basis of how they redistribute their output illumination power and how large an area they measure light over at an instant. In general, sensing speed (in terms of number of measurements obtained per unit time) increases when illumination and imaging are less concentrated.

The algorithms presented in chapters 2 and 3 use conventional projector-camera systems that illuminate and image the entire scene at once. They rely on computational techniques and carefully designed light patterns to handle the effects of global light transport, defocus blur and motion but they are not robust to ambient illumination. Concentrating light is crucial to building active illumination devices that work at high ambient light levels and it also makes handling the effects of global light transport easier.

Scanning laser rangefinders take this idea of robustness through spatial and temporal concentration to its logical extreme. They emit light in short, highly directional bursts. The light sensors they use are also highly directional in their sensitivity and are collocated to the source. Since the emitted light is concentrated both spatially and temporally, its reflection can be detected robustly by the sensor even at very high ambient light levels. Also, since most global light transport is fairly diffuse and only a single point is being illuminated at a time, very little global light reaches the collocated light sensor. The disadvantage of this scanning approach is that it is slow. The round-trip time of light places a hard limit on the rate at which points can be measured and in practice the need to physically rotate the source and sensor to change scanning direction slows down such systems further. Handling dynamic scenes is difficult and a separate sensing modality is often needed to track motion of the rangefinder so that point measurements can be fused together into a registered point cloud from a moving platform.

Flash lidars address the speed limitation and the need for moving parts. They emit short, undirected flashes of light and measure the round-trip time for reflected light with a two dimensional imaging array. These devices work well in sunlight. However, the difficulties in making an

	Point Illumination	Line Illumination	Area Illumination
Point Imaging	Scanning Lidar		
Line Imaging		Episcan3D 4 Epitof 5	
Area Imaging	MC3D [13]	Line Striping	Structured Light CW-ToF Flash Lidar

Table 6.1: Taxonomy of Active Illumination Sensors: Active illumination sensors can be classified on the basis of how the spatially concentrate illumination and imaging. Sensing speed in terms of measurements obtained per second typically increase as we move towards the bottom right corner where large parts of the scene are illuminated and imaged in parallel. Robustness to ambient illumination and global light transport affects can be achieved through spatial concentration (towards the top left corner in the table). Flash lidar systems are robust to ambient and global illumination despite using full-field illumination and sensing because they use highly temporally concentrated light pulses but are difficult to build at high resolutions. Episcan3D and Epitof strike a balance between full-field and point-wise sensing by sensing one line at a time. This provides reasonable robustness to ambient and global illumination without sacrificing measurement rate.

imaging array of sensitive photodetectors with the required temporal resolution for light timing measurements makes these devices expensive and they tend to have low spatial resolutions.

Structured light and continuous wave ToF [90] devices have typically not redistributed light spatially or temporally and so their performance degrades rapidly as ambient light levels increase (the suppression of background illumination [53] implemented on PMD sensors prevents ambient light from saturating the sensor but does nothing to solve the more fundamental problem of shot noise). These devices can achieve some robustness to global light transport effects through a combination of carefully designed illumination patterns [12, 26] or by acquiring extra measurements and computationally inverting some simplified model of global light transport [17, 18, 97]. Line striping devices [65] are an exception - they are structured light devices that concentrate their output light but they require one readout of the entire camera sensor for each line position so they are typically very slow.

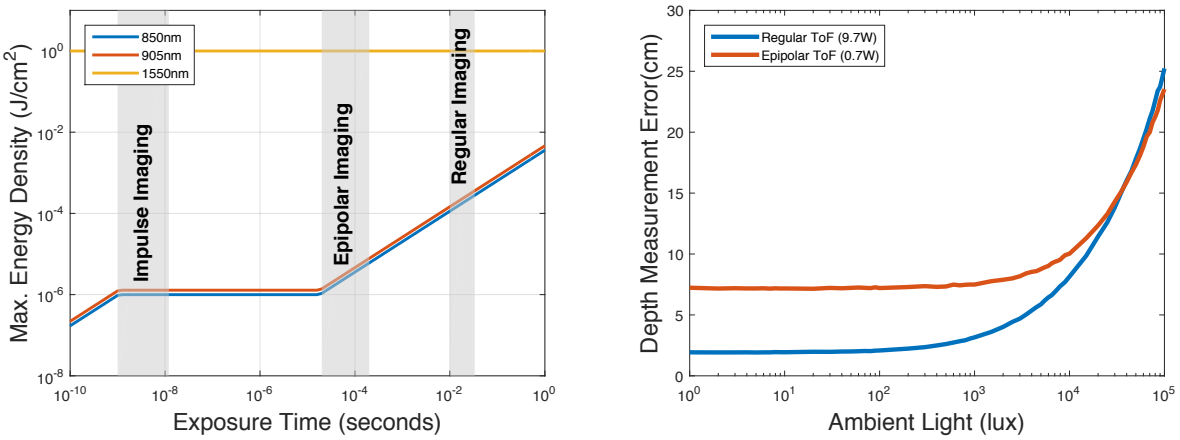
The epipolar imaging devices developed in this thesis – Episcan3D (chapter 4) and Epipolar Time-of-Flight (chapter 5) can be thought of as a compromise between the two extremes of full-field capture and point-by-point scanning. Because epipolar imaging illuminates and captures a single line at a time, it allows a depth camera to have most of the robustness to both ambient light and global illumination that point scanning provides while still having a high measurement rate.

6.1 Redistributive Light Sources and Eye Safety

Episcan3D and Epipolar Time-of-Flight use projectors that redistribute their light output dynamically instead of illuminating the entire field-of-view at once. In the case of Episcan3D, the projector illuminates a single point at a time and raster scans and the projector used in the Epipolar ToF camera illuminates a line at a time and the line sweeps across the field-of-view. It should be noted that epipolar imaging does not require a point scanning light source - a line scanning one suffices. Our Episcan3D prototype uses a raster scanning projector only because it was a convenient way to generate patterns with off-the-shelf hardware.

Given a fixed energy budget per measurement, in the presence of ambient light, epipolar imaging with a redistributive light source will always be more accurate than regular imaging with a non-redistributive light source (see figures 4.12 and 5.6). However, from an eye safety perspective, for visible and near-infrared light, shorter light pulses (like those generated by a redistributive source) must have lower energy density than longer pulses (like those generated by a non-redistributive light source). This is illustrated in figure 6.1a which plots the maximum safe energy density per unit area of a light pulse as measured at the eye for different pulse lengths and wavelengths.

The illumination pulse length for an impulse imaging systems like lidar (flash lidar or scan-



(a) Allowable Energy Density

(b) Epipolar vs Regular ToF Imaging

Figure 6.1: (a) In the near-infrared band, the permissible energy density of a light source is a function of exposure time. As a result, for a given eye safety requirement, regular imaging with a non-redistributive light source can use much higher power than epipolar imaging with a line striping source. (b) compares the accuracy of two CW ToF systems with the safe eye safety characteristics - one using regular imaging and another using epipolar. The target distance is fixed at $10\,\mathrm{m}$ and ambient light levels are varied. Despite using roughly 14 time less energy, epipolar imaging would still be more accurate than regular imaging in bright sunlight. For short-wave infrared, the maximum allowable energy density is much higher and doesn't depend on exposure time, but SWIR sensors are expensive and difficult to fabricate.

ning) or a high-speed raster scanning projector is usually in the $1\,\mathrm{ns}$ to $10\,\mathrm{ns}$ range. For a regular imaging system that doesn't use a redistributive light source, the illumination usually remains on for the entire frame exposure time, so typical pulse lengths would be between $10\,\mathrm{ms}$ and $30\,\mathrm{ms}$ for a system operating at video frame rates. For the epipolar imaging systems we've developed, typical illumination times per line for structured light or CW ToF are between $20\,\mu\mathrm{s}$ and $200\,\mu\mathrm{s}$. This places epipolar imaging just after the knee in the curve - a regular imaging system could operate at considerably higher total light output level and still be eye safe while an impulse imaging system could output roughly the same light energy as epipolar imaging but over a much shorter period.

One solution to this problem would be to operate in the short-wave infrared band (for example at $1550\,\mathrm{nm}$ instead of visible light or near-infrared. In this band, the permissible energy limits are much higher and also do not depend on pulse duration. Many lidar systems operate in the short-wave infrared band for this reason. Unfortunately, the technology for short-wave infrared sensors isn't yet as well developed as for sensors at other wavelengths. Another solution is to

use a multiple point sources or an extended light source as the permissible energy limits for such configurations are considerably higher. Many regular CW ToF devices use multiple sources or extended sources to increase the amount of power they can safely emit. In an epipolar imaging system this could be achieved by placing multiple low-power light sources along the epipolar line or using a VCSEL array.

Our Epipolar ToF prototype uses a $100\,\mu\mathrm{s}$ exposure per row, a $700\,\mathrm{mW}$ point source and is eye safe at $40\,\mathrm{cm}$ and beyond. A regular imaging equivalent would use a $24\,\mathrm{ms}$ exposure per frame (the sensor array has 240 rows) and would be eye safe at $40\,\mathrm{cm}$ even if it used a light source with a peak output power as high as $9.7\,\mathrm{W}$. Figure 6.1b plots how accurate depth measurements would be from these two devices under varying ambient light levels. Despite the fact that epipolar ToF is using almost 14 times less energy per frame than regular imaging, it performs just as well in bright sunlight.

6.2 What Sensor To Use Where?

Previously, sensor choice was largely dictated by environmental factors. Outdoors, passive stereo or lidar (either flash or scanning) were usually the only viable options. At short ranges, if power wasn't a major concern, continuous wave time-of-flight with high-powered illuminators (in the $10\,\mathrm{W}$ range), could also be used.

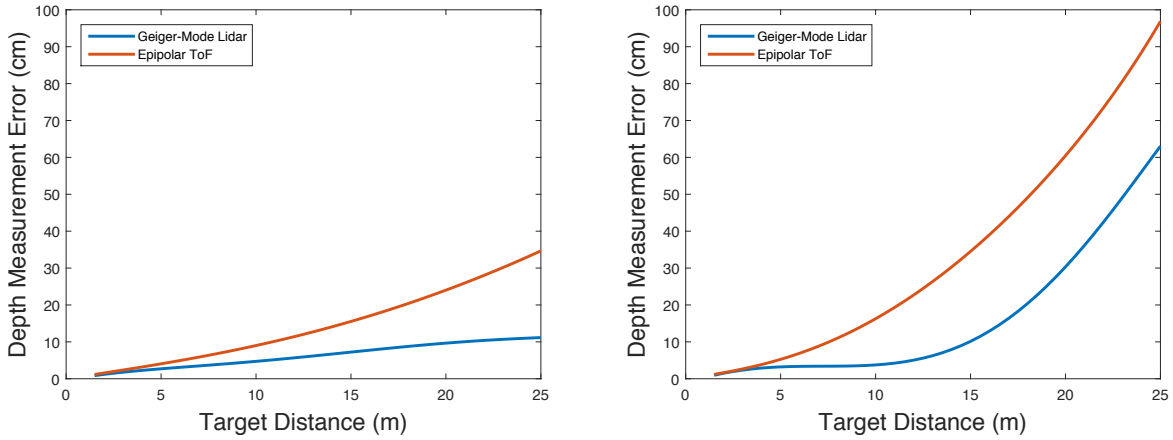
With epipolar imaging, structured light, active stereo and CW-ToF can be made to work outdoors, even in bright sunlight. This is significant because these active illumination methods can be realized with commodity hardware.

Epipolar imaging devices perform well at short to moderate ranges. But for long range sensing, lidar would still be the sensor of choice. Compared to Lidars, the range of an epipolar continuous wave time-of-flight system is limited by two factors:

- Sensitivity: Compared to the avalanche photo diodes (both linear and Geiger-mode) used by lidars, CW ToF sensors have low sensitivity - the photon flux required to generate a usable signal is high. Since strength of return falls of with the square of distance, low sensitivity reduces range.
- Ambient Light: Epipolar CW-ToF uses a far shorter exposure than regular CW-ToF systems, but the exposure time is still in the tens to hundreds of microseconds range. In comparison, a lidar needs to expose its sensor for only the round-trip time of the emitted light pulses which works out to $1\,\mu\mathrm{s}$ for a target $150\,\mathrm{m}$ away. Since CW-ToF sensors rely on integrated intensity readings to compute depth, the long exposure time results in increased shot noise due to ambient light which reduces accuracy.

Figure 6.2 compares the accuracy of Epipolar CW-ToF to Geiger-mode lidar. Using the same

amount of emitted energy per depth frame, Geiger-Mode lidar is more accurate than epipolar ToF. However, the accuracy of epipolar ToF is still adequate for many tasks and could still prove to be an attractive alternative to more complex and expensive lidar systems in some applications.



(a) Ambient Light Level $10\,\mathrm{klx}$ **(b)** Ambient Light Level $100\,\mathrm{klx}$

Figure 6.2: Simulated Comparisons of Epipolar CW-ToF to Geiger-Mode Lidar. For parameters of the simulated Epipolar CW-ToF camera see table 5.2. The Geiger-Mode lidar light source emits 100 pulses per depth reading, with a pulse length of $5\,\mathrm{ns}$ and per-pulse energy of $288\,\mu\mathrm{J}$. This makes the energy emitted per depth frame equal for both sensing modalities. Field of view, operating wavelength, sensor size and other parameters were also kept equal in both simulated systems. For the simulated Geiger-Mode sensor, the photon detection probability is 0.1, the detection jitter is $100\,\mathrm{ps}$ and the lens aperture is set to $f/4$. (a) compares the two systems under the equivalent of cloudy conditions, (b) compares the two systems under bright sunlight.

Chapter 7

Conclusion

Throughout this thesis we have worked towards the goal of creating active illumination systems that are robust to a variety of challenges present in uncontrolled, real-world environments and that can recover depth and capture components of light transport. These challenges include bright ambient light, interference from other active illumination systems, global light transport, defocus blur, and motion. Our aim has been to use the types of sensors and light sources that are used in consumer-grade depth cameras and build systems that have the robustness of a scanning lidar without sacrificing the compactness, economy and measurement density of a consumer depth camera.

Chapters 2 and 3 extended the capabilities of regular projector-camera systems that use non-redistributive light sources. They captured images that had been affected by indirect illumination, scene motion or illumination defocus and used computational models to try recover corrected images like those that would have been captured if there was no indirect illumination, the scene was stationary or there was no defocus blur.

In chapters 4 and 5 we adopt a markedly different approach. We built imaging systems that capture images that contain only the light paths of interest and block all (or most) of the unwanted light from confounding sources (ambient light, indirect light paths, interfering light from other systems). One of the keys to realizing such active imaging system was the use of redistributive light sources like raster scanning laser projectors and light sheet projectors.

7.1 Future Work

There are a number of ways that the ideas presented in this thesis could be developed further. We have developed two devices that implement epipolar imaging, one for structured light and another for continuous wave time-of-flight but there are many other hardware embodiments for the technique. On the computational side, there is work still to be done to address some of the

limitations of epipolar imaging when handling scene motion and certain types of global light transport.

Modeling Epipolar Indirect Light Paths Epipolar imaging prevents most (but not all) indirect light paths from reaching the sensor. There are two particular cases where global light transport can induce significant errors even with epipolar imaging. The first is specular interreflections that happen to line up with the epipolar lines (see bottom row of Figure 4.5 for an example). The second is when light scatters in a participating medium like fog.

Even when indirect light is present in epipolar imaging, it is much simpler to model than in the regular imaging case. For instance, specular interreflections are constrained to be caused by scene points that lie along the same epipolar line. In a participating medium, epipolar imaging will block most multiple-scattered light. Single-scattered light will reach the sensor as it obeys epipolar geometry, but single-scattering is much simpler than multiple scattering and can be modeled analytically [98].

What this means is that epipolar imaging could be combined with methods that attempt to handle indirect illumination by recovering transient responses [18, 40, 97]. The transient responses under epipolar imaging would be much sparser and have a higher signal (direct light path) to noise (indirect light path) ratio. This would allow recovering geometry with fewer measurements and in more challenging scenes.

Motion Registration and Correction Epipolar imaging allows for measurement sampling schemes like those presented in Figure 5.3 where blur and registration error-free depth measurements can be obtained even for rapid scene or camera motion. However, since each row of depth measurements in the image is captured at a different time the resulting depthmaps are distorted with an effect similar to rolling-shutter. Tailoring methods that can correct these types of distortions such as [4, 49] to work with epipolar imaging systems is an important avenue for future work.

Implementing Epipolar Imaging With a Line Sensor In this thesis we implemented sensor-side masking for epipolar imaging via two mechanisms - exploiting the rolling shutter in the case of Episcan3D and controlling the region of interest (ROI) in the Epitof system.

In both cases, we use sensors that have a 2D array of pixels, but only a single row of pixels in the sensor is active at a time. In some sense, almost all the pixels are being 'wasted' at any given point of time. For regular CMOS sensors, this is not a major factor because even large CMOS pixel arrays are cheap and easy to fabricate. But for sensors with more complex pixel designs (like short wave infrared sensors, Dynamic Vision Sensors [61] or a flash lidars) building large

2D pixel arrays can be both difficult and expensive. In these cases an epipolar imaging device design that required a sensor with just a single line of pixels as opposed to a 2D array of pixels would be desirable.

This sort of line sensor based epipolar imaging system could be realized by placing a controllable 1D scanning mirror in front of a line sensor camera. Steering the scanning mirror would select the plane in the world along which the. One important consideration in such an implementation would be the design of the line sensor's lens and the placement and size of the scanning mirror. To prevent the mirror from being a bottleneck in terms of light throughput, it would need to be large enough to cover the entire aperture of the lens. However, a very large mirror is undesirable because it would be bulky, scan slowly and require a lot of power to drive. The imaging lens needs to be designed so that its aperture is as close as possible to the front element. This would allow the imaging scanning mirror to be small while not restricting light throughput. The system could also potentially use the same scanning mirror for illumination and imaging.

Another advantage of implementing epipolar imaging with a line sensor is that since the photosensitive elements are arranged in a line, the space above and below the photosensors can be used for frame transfer and readout electronics which leads to higher fill factors than in a 2D pixel array.

Photometric Stereo Photometric stereo is very sensitive to global light transport effects. For instance, interreflections cause concavities to systematically appear shallower than they actually are when reconstructed using photometric stereo [25]. Photometric stereo requires images captured with at least three different lighting directions. If there were a means to capture epipolar-only images with illumination from different directions, they could be used to perform more accurate photometric stereo.

The Episcan3D imaging system can easily be extended to two lighting directions by adding a second projector light source along the original projector-camera baseline axis, but on the other side of the camera. With such an arrangement, epipolar-only images with two different lighting directions can be captured. Adding more projectors along the same axis doesn't help because lighting directions that are contained by a plane form a degenerate case.

Instead, a third (and even a forth) light source can be placed along the axis perpendicular to the system's original baseline. The third (and fourth) projector would be a rectified configuration with respect to the camera, the only difference being that they would be rectified along vertically (along pixel columns) instead of horizontally along rows. This configuration would be sufficient to recover epipolar-only images under four different lighting directions for photometric stereo. Unfortunately, with the rolling-shutter masks used by Episcan3D, there is no way to mask pixels along both rows (for projectors one and two) and columns (for projectors three and

four) in one camera. However, a sensor that supports almost arbitrary pixel masks was recently developed [86]. With such a sensor, epipolar photometric stereo could be implemented.

More Selective Imaging Both the epipolar imaging devices we have built expose one row of pixels at a time. With more fine-grained control over camera pixel masking, it would be possible to implement variants of epipolar imaging that are even more selective. For instance, in the Episcan3D system only a short subsegment of a row could be exposed at a time and the exposed pixels could be synchronized to the raster-scan path of the projector. This would make it possible to use shorter exposure times per pixel (resulting in even less ambient light reaching the sensor) and also combine epipolar structured light with depth gating (resulting in less global illumination effects). It maybe possible to realize such types of masking with the a sensor like the one developed in [86].

Appendix

This appendix provides Matlab code listings for the functions and scripts used to generate the simulation results presented in this thesis for Episcan and Epipolar Time-of-Flight.

Simulation Code for Episcan3D

Listing 1: generateFigure4point12.m

```
snr = 3;                %target SNR

%scene specs
params.La     = 1e4;
params.albedo = 0.5;
params.rho    = params.albedo;

%projector specs
%operating wavelength
params.wavelength = 532e-9;
%projector power at operating wavelength in lumens
proj_lumens = 10 ;
%projector area at u throw distance
area1u      = 0.5;
params.u    = 1   ; % meters
%Projector lux at distance u
params.Lp1 = proj_lumens/area1u;

%camera specs
params.pixel_side   = 5e-6;
params.pixel_area   = params.pixel_side^2;
params.quantum_eff  = 0.7;
params.filter_trans = 0.95;
```

```
24      params.focal_length = 5e-3;
25      params.fnumber      = 1.0/1.6;
26      %ambient blocking factor
27      params.k            = 1/500;
28      params.exposure_time= 1/60;
29
30      params.max_electrons=12e3;
31      params.read_noise_electrons=10;
32
33      params0 = params;
34
35      params_all = cell(4,1);
36
37      %no filter, no suppression
38      params_all{1}              = params0;
39      params_all{1}.filter_trans = 1;
40      params_all{1}.k            = 1; %no blocking
41
42      % filter, no suppression
43      params_all{2}              = params0;
44      params_all{2}.filter_trans = 1;
45      % ambient blocking factor of 15 from filter
46      params_all{2}.k            = 1/15;
47
48      % filter, yes suppression
49      params_all{3}              = params0;
50      params_all{3}.filter_trans = 1;
51      % 15 from filter, 160 from epipolar imaging
52      params_all{3}.k            = 1/15 * 1/160;
53
54      % filter, ideal suppression
55      params_all{4}              = params0;
56      params_all{4}.filter_trans = 1;
57      % 15 from filter, 800 from episcan
58      params_all{4}.k            = 1/15 * 1/800;
59
60      La_test = 10.^(0:0.25:5);
61      z_all = zeros(numel(params_all), numel(La_test));
62
```

```
63    for i=1:numel(params_all)
64        for j=1:numel(La_test)
65            tmp = params_all{i};
66            tmp.La = La_test(j);
67            z_all(i,j) = fminsearch(...
68                @(z)simulateMaxRangeResidue(z, snr, tmp), 0);
69        end
70    end
71
72    figure(1);
73    semilogx(La_test, z_all, 'linewidth', 3);
74    xlabel('Ambient Illumination (lux)');
75    ylabel('Working Range (meters)');
76    axis([1 10^5 0 3]);
77
78    ztest = 5;
79
80    power_all = zeros(numel(params_all), numel(La_test));
81
82    for i=1:numel(params_all)
83        for j=1:numel(La_test)
84            tmp = params_all{i};
85            tmp.La = La_test(j);
86            power_all(i,j) = ...
87              fminsearch(@(L)simulateMinLpResidue(...
88                L, ztest, snr, tmp), 1000);
89        end
90    end
91
92    figure(2);
93    semilogx(La_test, power_all, 'linewidth', 3);
94    xlabel('Ambient Illumination (lux)');
95    ylabel('Minimum Projector Power (lumens)');
96    axis([1 1e5 0 1000]);
```

Listing 2: simulateMaxRangeResidue.m

```
function res = simulateMaxRangeResidue(z, snr, params)
% Returns 0 if the computed SNR at depth z matches snr.

[mu1, sigma1] = simulatePhotons(1,z,params);
[mu0, sigma0] = simulatePhotons(0,z,params);

mu = mu1 - mu0;
sigma = sigma0 + sigma1;

res = (mu - snr*sigma)^2;
```

Listing 3: simulateMinLpResidue.m

```
function res = simulateMinLpResidue(Lp, z, lambda, params)

params.Lp1 = Lp;
[mu1, sigma1] = simulatePhotons(1,z,params);
[mu0, sigma0] = simulatePhotons(0,z,params);

mu = mu1 - mu0;
sigma = sigma0 + sigma1;
res = (mu - lambda*sigma)^2;
```

Listing 4: simulatePhotons.m

```
function [mu, sigma] = simulatePhotons(projector_val, z, params)

planks        = 6.626e-34;
lightSpeed    = 3e8;
lumen_to_watt = 1/683;
photon_energy = planks*lightSpeed ...
                /params.wavelength;


L = params.rho * (params.k * params.La + ...
    projector_val * params.filter_trans * ...
    (params.u^2)/(z^2)*params.Lp1);


P = params.quantum_eff * L * ...
    params.pixel_area/(2*pi) * ...
    params.fnumber^2 * lumen_to_watt;
E = P * params.exposure_time;

mu = E / photon_energy;
sigma = sqrt(mu) + params.read_noise_electrons;
```

Simulation Code for Epipolar Time-of-Flight Imaging.

Listing 5: generateFigure5point6.m

```
%% noise in estimated depth at 10m

epi_projector_power = 2; %watts
reg_projector_power = 2; %watts

scene.z = 10;
scene.albedo = 0.5;

ambient_levels = [0 (1:9)*1e-5 ...
                  (1:9)*1e-4 (1:9)*1e-3 ...
                  (1:0.5:9)*1e-2 (1:0.5:9)*1e-1 1];

epi_camera   =  defaultCamera();
epi_camera.exposure = 30e-6;           %seconds
epi_projector = defaultProjector();
%watts per pixel
epi_projector.power_per_pixel = epi_projector_power/320;

reg_camera   =  defaultCamera();
reg_camera.exposure = 240*30e-6;           %seconds
reg_projector = defaultProjector();
%watts per pixel
reg_projector.power_per_pixel = reg_projector_power/(320*240);

epi_std = zeros(size(ambient_levels));
reg_std = zeros(size(ambient_levels));

for i=1:numel(ambient_levels)
    [~, epi_var, ~] = computeToFPhaseStatistics(...
        scene, epi_camera, epi_projector, ...
        ambient_levels(i), 20000);
    epi_std(i) = sqrt(epi_var);
    [~, reg_var, ~] = computeToFPhaseStatistics(...
        scene, reg_camera, reg_projector, ...
        ambient_levels(i), 20000);
```

```
36          reg_std(i) = sqrt(reg_var);
37      end
38
39      % Convert standard deviation from phase to z error.
40      epi_std = epi_std/(2*pi)*3e8/epi_projector.modulation_freq;
41      reg_std = reg_std/(2*pi)*3e8/epi_projector.modulation_freq;
42
43      figure(2);
44      hold off;
45      semilogx(1e5*ambient_levels(2:end), ...
46              100*reg_std(2:end),'linewidth',4);
47      hold on;
48      semilogx(1e5*ambient_levels(2:end), ...
49              100*epi_std(2:end),'linewidth',4);
50      axis([0 1e5 0 30]);
51
52      %legend('Regular ToF', 'Epipolar ToF', 'Location', 'NorthWest');
53      %xlabel('Ambient Light (lux)','FontSize',16);
54      %ylabel('Depth Measurement Error(cm)','FontSize',16);
55
56      % Working range as function of desired range accuracy.
57
58      epi_camera    =  defaultCamera();
59      epi_camera.exposure = 30e-6;           %seconds
60      epi_projector = defaultProjector();
61      %watts per pixel
62      epi_projector.power_per_pixel = projector_power/320;
63
64      reg_camera    =  defaultCamera();
65      reg_camera.exposure = 240*30e-6;           %seconds
66      reg_projector = defaultProjector();
67      %watts per pixel
68      reg_projector.power_per_pixel = projector_power/(320*240);
69
70      lambda = 3e8/epi_projector.modulation_freq;
71
72      target_std = 0.05; %meters
73      target_std = target_std/lambda * (2*pi); %radians
74
```

```
75      z_epi = zeros(size(ambient_levels));
76      z_reg = zeros(size(ambient_levels));
77
78      parfor i=1:numel(ambient_levels)
79          zmin = 1;
80          zmax = 20;
81          zcur = 0;
82          for j=1:11
83              nSamples = 2000 * (1+j);
84
85              zcur = (zmin + zmax)/2;
86              scene = [];
87              scene.albedo = 0.5;
88              scene.z = zcur;
89              [~, epi_var, ~] = computeToFPhaseStatistics(...
90                scene, epi_camera, epi_projector, ...
91                ambient_levels(i), nSamples);
92              epi_std = sqrt(epi_var);
93              if (epi_std > target_std)
94                  zmax = zcur + (zmax-zmin)*0.1;
95              else
96                  zmin = max(0, zcur - (zmax-zmin)*0.1);
97              end
98          end
99          z_epi(i) = zcur;
100
101         zmin = 1;
102         zmax = 15;
103         for j=1:11
104             nSamples = 2000 * (1+j);
105
106             zcur = (zmin + zmax)/2;
107             scene = [];
108             scene.albedo = 0.5;
109             scene.z = zcur;
110             [~, reg_var, ~] = computeToFPhaseStatistics(...
111               scene, reg_camera, reg_projector, ...
112               ambient_levels(i), nSamples);
113             reg_std = sqrt(reg_var);
```

```
114             if (reg_std > target_std)
115                 zmax = zcur + (zmax-zmin)*0.1;
116             else
117                 zmin = max(0, zcur - (zmax-zmin)*0.1);
118             end
119         end
120         z_reg(i) = zcur;
121     end
122
123
124     figure(3);
125     hold off;
126     semilogx(1e5*ambient_levels(2:end), z_reg(2:end),'linewidth',3);
127     hold on;
128     semilogx(1e5*ambient_levels(2:end), z_epi(2:end),'r','linewidth',3);
129     legend('Regular ToF', 'Epipolar ToF', 'Location', 'SouthWest');
130     xlabel('Ambient Illumination (lux)');
131     ylabel('Working Range (m)');
```

Listing 6: defaultToFCamera.m

```
function camera = defaultToFCamera()

%camera settings
camera.exposure           = 100e-6; %in seconds

%camera physical properties
camera.focal_length       = 8e-3 ; %focal length (in meters)
camera.fnumber            = 1.1  ; %aperture fnumber

%camera pixel parameters
camera.pixel_side         = 20e-6; %pixel side (in meters)
camera.quantum_efficiency = 0.8;
camera.pixel_area         = camera.pixel_side^2;

%bandpass filter parameters
camera.filter_center    = 850e-9; %filter center (in meters)
camera.filter_bandwidth =  20e-9; %filter bandwidth (in meters)

%noise and digitization
camera.read_noise       = 5; %(electrons per readout)
camera.electrons_per_LSB= 1;

%spad only parameters
camera.detection_probability = 0.1;
camera.detection_jitter = 40e-12; % seconds
```

Listing 7: defaultToFProjector.m

```
function projector = defaultToFProjector()

projector.power_per_pixel = 700e-3 / 320; %Watts per pixel
projector.modulation_freq = 15e6;         %Hz
```

Listing 8: computePhotonArrivalRate.m

```
function lambda = computePhotonArrivalRate(
                        scene, camera, projector, ...
                        relative_projector, relative_ambient)

% mu is the detection rate at pixel in electrons per second.
% relative_projector is a number between 0 and 1.
% relative_ambient is a number between 0 and 1.

ambient_spectral_density = 0.85e9; % W/m
if (camera.filter_center < 800e-9)
    ambient_spectral_density = 1.25e9; % W/m
end

scene_patch_area   = (scene.z/camera.focal_length)^2 ...
                     * camera.pixel_area;
ambient_at_patch   = relative_ambient * ...
                     ambient_spectral_density * ...
                     camera.filter_bandwidth * ...
                     scene_patch_area;
power_at_patch     = ambient_at_patch + ...
                     relative_projector * ...
                     projector.power_per_pixel;

aperture_diameter = camera.focal_length / camera.fnumber;
aperture_area     = pi/4 * aperture_diameter^2;
subtended_angle   = aperture_area / (4*pi*scene.z^2);

incident_power    = scene.albedo * power_at_patch * ...
                    subtended_angle / (2*pi);

energy_per_photon =  6.626e-34*3e8/camera.filter_center;
incident_photons  = incident_power / energy_per_photon;
lambda            = incident_photons *  camera.quantum_efficiency;
```

Listing 9: computeExpectedQuadratureReading.m

```
function mu = computeExpectedQuadratureReading(...
                          scene, camera, projector, relative_ambient)

assert(nargin == 4);

c = 3e8; % m/s
phase = 4*pi*scene.z*projector.modulation_freq/c;

%lambda is the electron arrival rate at each quadrature reading
lambda = zeros(4,1);
lambda(1) = computePhotonArrivalRate(...
        scene, camera, projector, ...
        0.5*(1+cos(phase)), relative_ambient);
lambda(2) = computePhotonArrivalRate(...
        scene, camera, projector, ...
        0.5*(1+sin(phase)), relative_ambient);
lambda(3) = computePhotonArrivalRate(...
        scene, camera, projector, ...
        0.5*(1-cos(phase)), relative_ambient);
lambda(4) = computePhotonArrivalRate(...
        scene, camera, projector, ...
        0.5*(1-sin(phase)), relative_ambient);

%mu is the mean value in electrons for each quadrature reading

mu = lambda * camera.exposure + camera.read_noise;
```

Listing 10: computeToFPhaseStatistics.m

```
function [mu, variance, deciles] =  ...
    computeToFPhaseStatistics(scene, camera, projector, ...
                              relative_ambient, nsamples)

if (nargin < 5)
    nsamples = 1e4;
end

c = 3e8; % m/s

phase_true    = 4*pi*scene.z*projector.modulation_freq/c;
phase_samples = zeros(nsamples,1);

mean_reading  = computeExpectedQuadratureReading(...
                    scene,camera, projector,relative_ambient);

for i=1:nsamples
    reading = poissrnd(mean_reading);
    reading = round(reading/camera.electrons_per_LSB);
    phase_samples(i) = atan2(reading(2)-reading(4), ...
                             reading(1)-reading(3));
end

phase_samples = nAngle(phase_samples);
phase_samples(phase_samples - phase_true > pi) = ...
phase_samples(phase_samples - phase_true > pi) - 2*pi;
phase_samples(phase_samples - phase_true <-pi) = ...
phase_samples(phase_samples - phase_true <-pi) + 2*pi;

mu       = angularMean(phase_samples);
if (phase_true - mu > pi); mu = mu + 2*pi; end;
if (phase_true - mu <-pi); mu = mu - 2*pi; end;

variance = angularVariance(phase_samples);
deciles  = quantile(phase_samples, 0:0.1:1);
```

Bibliography

[1] Supreeth Achar and Srinivasa G Narasimhan. Multi focus structured light for recovering scene shape and global illumination. In *Computer Vision–ECCV 2014*, pages 205–219. Springer, 2014. 1.4

[2] Supreeth Achar, Stephen T Nuske, and Srinivasa G Narasimhan. Compensating for Motion During Direct-Global Separation. *International Conference on Computer Vision*, 2013. 1.4

[3] Supreeth Achar, Joseph R. Bartels, William L. Whittaker, Kiriakos N. Kutulakos, and Srinivasa G. Narasimhan. Epipolar time-of-flight imaging. *ACM Trans. Graph.*, 36(4), July 2017. ISSN 0730-0301. URL `http://dx.doi.org/10.1145/3082959.3073686`. 1.4

[4] Hatem Alismail, L Douglas Baker, and Brett Browning. Continuous trajectory estimation for 3d slam from actuated lidar. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6096–6101. IEEE, 2014. 5.3.1, 7.1

[5] American National Standards Institute. *American National Standard for Safe Use of Lasers Z136.1*. Laser Institute of America, 2014. ISBN 9781940168005. 5.4.4

[6] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *TPAMI*, 2015. 1

[7] EV Browell, S Ismail, and WB Grant. Differential absorption lidar (dial) measurements from air and space. *Applied Physics B: Lasers and Optics*, 67(4):399–410, 1998. 1.1

[8] Thomas Brox, Nils Papenberg, and Joachim Weickert. High Accuracy Optical Flow Estimation Based on a Theory for Warping. *European Conference on Computer Vision*, 4 (May):25–36, 2004. 2.3.2

[9] Bernhard Buttgen and Peter Seitz. Robust optical time-of-flight range imaging based on smart pixel structures. *IEEE Trans. on Circuits and Systems*, 55-I(6):1512–1525, 2008. 5.1

[10] Antonin Chambolle. An Algorithm for Total Variation Minimization and Applications. *Journal of Mathematical Imaging and Vision*, 20(1-2):89–97, 2004. 2.3.1

[11] ML Chanin, A Garnier, A Hauchecorne, and J Porteneuve. A doppler lidar for measuring winds in the middle atmosphere. *Geophysical research letters*, 16(11):1273–1276, 1989. 1.1

[12] Tongbo Chen, Hans-Peter Seidel, and Hendrik P. Lensch. Modulated phase-shifting for

3D scanning. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008. doi: 10.1109/CVPR.2008.4587836. 3.1.1, 6

[13] Oliver S Cossairt, Nathan Matsuda, and Mohit Gupta. Motion contrast 3d scanning. In *Computational Optical Sensing and Imaging*, pages CT2E–1. Optical Society of America, 2015. 1.3.3, 5.1, **??**

[14] Vincent Couture, Nicolas Martin, and Sebastien Roy. Unstructured light scanning to overcome interreflections. *International Conference on Computer Vision*, 2011. doi: 10.1109/ICCV.2011.6126458. 1.3.1, 2.3.1, 3.1.1

[15] James Davis, Ravi Ramamoorthi, and Szymon Rusinkiewicz. Spacetime stereo: A unifying framework for depth from triangulation. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–359. IEEE, 2003. 1.3.5

[16] B. DE Cremoux. Avalanche photo-diodes, May 25 1976. URL `https://www.google.com/patents/US3959646`. US Patent 3,959,646. 1.2.1

[17] Eric Dedrick. Improving sli performance in optically challenging environments. 2011. 6

[18] Adrian A Dorrington, John Peter Godbaz, Michael J Cree, Andrew D Payne, and Lee V Streeter. Separating true range measurements from multi-path and scattering interference in commercial range cameras. In *IS&T/SPIE Electronic Imaging*, pages 786404–786404. International Society for Optics and Photonics, 2011. 5.2, 6, 7.1

[19] C Dunsby and P M W French. Techniques for depth-resolved imaging through turbid media including coherence-gated imaging. *Journal of Physics D: Applied Physics*, 36(14): R207, 2003. URL `http://stacks.iop.org/0022-3727/36/i=14/a=201`. 1.1

[20] Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1033–1038. IEEE, 1999. 2.3.1

[21] Michael Elad, J-L Starck, Philippe Querre, and David L Donoho. Simultaneous cartoon and texture image inpainting using morphological component analysis (mca). *Applied and Computational Harmonic Analysis*, 19(3):340–358, 2005. 2.3.1

[22] Stefan Fuchs. Multipath interference compensation in time-of-flight camera images. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3583–3586. IEEE, 2010. 1.3.1

[23] Luis Goddyn, George M Lawrence, and Evi Nemeth. Gray codes with optimized run lengths. *Utilitas Mathematica*, 34:179–192, 1988. 4.4.2

[24] Dan B Goldman, Brian Curless, Aaron Hertzmann, and Steven M Seitz. Shape and spatially-varying BRDFs from photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):1060–71, June 2010. ISSN 1939-3539. doi: 10.1109/TPAMI.2009.102. 1.1

[25] Jinwei Gu, Toshihiro Kobayashi, Mohit Gupta, and Shree K. Nayar. Multiplexed illumination for scene recovery in the presence of global illumination. *International Conference*

on Computer Vision, pages 691–698, November 2011. doi: 10.1109/ICCV.2011.6126305. 1.3.1, 7.1

[26] Mohit Gupta and Shree K Nayar. Micro Phase Shifting. *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 3.1, 3.1.1, 3.6, 3.6.1, 4.4.2, 6

[27] Mohit Gupta, Amit Agrawal, Ashok Veeraraghavan, and Srinivasa G. Narasimhan. Structured light 3D scanning in the presence of global illumination. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 713–720, June 2011. doi: 10.1109/CVPR.2011.5995321. 1.3.1, 3.1.1, 4.4.2

[28] Mohit Gupta, Yuandong Tian, Srinivasa G. Narasimhan, and Li Zhang. A Combined Theory of Defocused Illumination and Global Light Transport. *International Journal of Computer Vision*, 98(2):146–167, October 2011. ISSN 0920-5691. doi: 10.1007/s11263-011-0500-9. 1.3.2

[29] Mohit Gupta, Yuandong Tian, Srinivasa G. Narasimhan, and Li Zhang. A Combined Theory of Defocused Illumination and Global Light Transport. *International Journal of Computer Vision*, 98(2):146–167, October 2011. ISSN 0920-5691. doi: 10.1007/s11263-011-0500-9. 3.1.1, 3.6.2

[30] Mohit Gupta, Qi Yin, and Shree K Nayar. Structured Light In Sunlight. *International Conference on Computer Vision*, 2013. 1.3.3, 5.1

[31] Mohit Gupta, Shree K. Nayar, Matthias B. Hullin, and Jaime Martín. Phasor imaging: A generalization of correlation-based time-of-flight imaging. *ACM Transactions on Graphics*, 34, 2015. 1.3.1, 5.2

[32] Otkrist Gupta, Thomas Willwacher, Andreas Velten, Ashok Veeraraghavan, and Ramesh Raskar. Reconstruction of hidden 3d shapes using diffuse reflections. *Opt. Express*, 20(17):19096–19108, Aug 2012. doi: 10.1364/OE.20.019096. URL `http://www.opticsexpress.org/abstract.cfm?URI=oe-20-17-19096`. 1.1

[33] D.S. Hall. High definition lidar system, 2011. URL `https://www.google.com/patents/US7969558`. US Patent 7,969,558. 1.2.1

[34] Kent Hansen, Jeppe Pedersen, Thomas Sølund, Henrik Aanæs, and Dirk Kraft. A structured light scanner for hyper flexible industrial automation. In *3D Vision (3DV), 2014 2nd International Conference on*, volume 1, pages 401–408. IEEE, 2014. 4.4.2

[35] Samuel W Hasinoff and Kiriakos N Kutulakos. Confocal stereo. In *Computer Vision–ECCV 2006*, pages 620–634. Springer, 2006. 1

[36] Samuel W Hasinoff and Kiriakos N Kutulakos. Light-efficient photography. *IEEE PAMI*, 33(11):2203–14, November 2011. ISSN 1939-3539. doi: 10.1109/TPAMI.2011.62. URL `http://www.ncbi.nlm.nih.gov/pubmed/21422486`. 2.5

[37] H.W. Haussecker and D.J. Fleet. Computing optical flow with physical models of brightness variation. *IEEE PAMI*, 23(6), June 2001. ISSN 01628828. doi: 10.1109/34.927465. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=927465`. 2.1.1

[38] Felix Heide, Matthias B Hullin, James Gregson, and Wolfgang Heidrich. Low-budget

transient imaging using photonic mixer devices. *ACM Transactions on Graphics (TOG)*, 32(4):45, 2013. 1.3.5

[39] Felix Heide, Lei Xiao, Wolfgang Heidrich, and Matthias B. Hullin. Diffuse mirrors: 3d reconstruction from diffuse indirect illumination using inexpensive time-of-flight sensors. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 3222–3229, Washington, DC, USA, 2014. IEEE Computer Society. ISBN 978-1-4799-5118-5. doi: 10.1109/CVPR.2014.418. URL http://dx.doi.org/10.1109/CVPR.2014.418. 1

[40] Felix Heide, Lei Xiao, Andreas Kolb, Matthias B. Hullin, and Wolfgang Heidrich. Imaging in scattering media using correlation image sensors and sparse convolutional coding. *Opt. Express*, 22(21):26338–26350, Oct 2014. doi: 10.1364/OE.22.026338. URL http://www.opticsexpress.org/abstract.cfm?URI=oe-22-21-26338. 5.2, 7.1

[41] Felix Heide, Wolfgang Heidrich, Matthias Hullin, and Gordon Wetzstein. Doppler time-of-flight imaging. *ACM Trans. Graph.*, 34(4):36:1–36:11, July 2015. ISSN 0730-0301. doi: 10.1145/2766953. URL http://doi.acm.org/10.1145/2766953. 1.1

[42] C. Hermans, Y. Francken, T. Cuypers, and P. Bekaert. Depth from sliding projections. *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. doi: 10.1109/CVPR.2009.5206610. 3.1

[43] Aaron Hertzmann and Steven M Seitz. Example-based photometric stereo: Shape reconstruction with general, varying brdfs. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1254–1264, 2005. 1.1

[44] B. K.P. Horn. Shape from shading: A method for obtaining the shape of a smooth opaque object from one view. Technical report, Cambridge, MA, USA, 1970. 1

[45] David Huang, Eric A Swanson, Charles P Lin, Joel S Schuman, William G Stinson, Warren Chang, Michael R Hee, Thomas Flotte, Kenton Gregory, Carmen A Puliafito, et al. Optical coherence tomography. *Science*, 254(5035):1178–1181, 1991. 1.1

[46] S Inokuchi, Sato K, and Matsuda F. Range imaging system for 3-d object recognition. In *International Conference on Pattern Recognition*, 1984. 1.3.5, 3.1

[47] Achuta Kadambi, Refael Whyte, Ayush Bhandari, Lee Streeter, Christopher Barsi, Adrian Dorrington, and Ramesh Raskar. Coded time of flight cameras: Sparse deconvolution to address multipath interference and recover time profiles. *ACM Trans. Graph.*, 32(6): 167:1–167:10, November 2013. ISSN 0730-0301. doi: 10.1145/2508363.2508428. URL http://doi.acm.org/10.1145/2508363.2508428. 5.2

[48] Hiroshi Kawasaki, Ryo Furukawa, Ryusuke Sagawa, and Yasushi Yagi. Dynamic scene shape reconstruction using a single structured light pattern. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008. doi: 10.1109/CVPR.2008.4587702. 1.3.5

[49] C. Kerl, J. Stueckler, and D. Cremers. Dense continuous-time tracking and mapping with rolling shutter RGB-D cameras. In *ICCV*, Santiago, Chile, 2015. 5.3.1, 7.1

[50] Leonid Keselman, John Iselin Woodfill, Anders Grunnet-Jepsen, and Achintya Bhowmik. Intel realsense stereoscopic depth cameras. *arXiv preprint arXiv:1705.05548*, 2017. 1.2.3

[51] Soren Konig and Stefan Gumhold. Image-Based Motion Compensation for Structured Light Scanning of Dynamic Surfaces. *Dynamic 3D Imaging Workshop*, 2007. 1.3.5, 2.1.1

[52] Kurt Konolige. Projected texture stereo. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 148–155. IEEE, 2010. 1.2.3

[53] H Kraft, J Frey, T Moeller, M Albrecht, M Grothof, B Schink, H Hess, and B Buxbaum. 3d-camera of high 3d-frame rate, depth-resolution and background light elimination based on improved pmd (photonic mixer device)-technologies. *OPTO, Nuernberg, May*, 2004. 6

[54] Bruce Lamond and Paul Debevec. Fast Image-based Separation of Diffuse and Specular Reflections. *SIGGRAPH Sketches*, pages 1–7, 2007. 1.1, 2.5

[55] Edwin H. Land and John J. McCann. Lightness and retinex theory. *J. Opt. Soc. Am.*, 61(1):1–11, Jan 1971. doi: 10.1364/JOSA.61.000001. URL `http://www.osapublishing.org/abstract.cfm?URI=josa-61-1-1`. 1

[56] Robert Lange and Peter Seitz. Solid-state time-of-flight range camera. *IEEE Journal of quantum electronics*, 37(3):390–397, 2001. 1.2.2, 5.2

[57] Thomas Laux. Asc's 3d flash lidar camera: The science behind asc's 3d depth imaging video camera. In *Stereoscopic 3D for Media and Entertainment, SMPTE International Conference on*, pages 1–10. SMPTE, 2010. 1.2.1

[58] Byron S. Lee and Timothy C. Strand. Profilometry with a coherence scanning microscope. *Appl. Opt.*, 29(26):3784–3788, Sep 1990. doi: 10.1364/AO.29.003784. URL `http://ao.osa.org/abstract.cfm?URI=ao-29-26-3784`. 1.1

[59] Seungkyu Lee, Byongmin Kang, James DK Kim, and Chang Yeong Kim. Motion blur-free time-of-flight range sensor. In *IS&T/SPIE Electronic Imaging*, pages 82980U–82980U. International Society for Optics and Photonics, 2012. 1.3.5

[60] Shuangyan Lei and Song Zhang. Digital sinusoidal fringe pattern generation: Defocusing binary patterns VS focusing sinusoidal patterns. *Optics and Lasers in Engineering*, 48(5): 561–569, May 2010. ISSN 01438166. doi: 10.1016/j.optlaseng.2009.12.002. 1.2.3, 3.1.1

[61] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128 by 128 120 db 15 microseconds latency asynchronous temporal contrast vision sensor. *IEEE journal of solid-state circuits*, 43(2):566–576, 2008. 7.1

[62] Satya P Mallick, Todd Zickler, Peter N Belhumeur, and David J Kriegman. Specularity Removal in Images and Videos : A PDE Approach. *European Conference on Computer Vision*, 2006. 2.5

[63] David Marr, G Palm, and T Poggio. Analysis of a cooperative stereo algorithm. *Biological Cybernetics*, 28(4):223–239, 1978. 1

[64] Paul F McManamon. *Field Guide to Lidar*. SPIE Press, 2015. 1.2.1

[65] Christoph Mertz, Sanjeev J. Koppal, Solomon Sia, and Srinivasa Narasimhan. A low-

power structured light sensor for outdoor scene reconstruction and dominant material identification. *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 15–22, June 2012. doi: 10.1109/CVPRW.2012.6239194. 1.3.3, 5.1, 6

[66] Yana Mileva. Illumination-Robust Variational Optical Flow with Photometric Invariants. *DAGM Conference on Pattern Recognition*, pages 152–162, 2007. 2.1.1

[67] M. Minsky. Microscopy apparatus, December 19 1961. URL `http://www.google.com/patents/US3013467`. US Patent 3,013,467. 1.1, 5.1

[68] Shree K Nayar, Masahiro Watanabe, and Minori Noguchi. Real-time focus range sensor. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(12):1186–1198, 1996. 1

[69] Shree K. Nayar, Gurunandan Krishnan, Michael D. Grossberg, and Ramesh Raskar. Fast separation of direct and global components of a scene using high frequency illumination. *ACM Transactions on Graphics*, 25(3):935, July 2006. ISSN 07300301. doi: 10.1145/1141911.1141977. 1.1, 1.3.1, 1.3.5, 1.4, 2.1, 2.1, 2.2, 2.3.3, 3.2, 3.7

[70] Shree K Nayar, Gurunandan Krishnan, Michael D Grossberg, and Ramesh Raskar. Visual chatter in the real world. In *Robotics Research*, pages 13–24. Springer, 2010. 2.8

[71] S.K. Nayar, K. Ikeuchi, and T. Kanade. Shape from Interreflections. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2–11, Dec 1990. 1.3.1

[72] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 343–352, 2015. 5.1

[73] Cristiano Niclass, Claudio Favi, Theo Kluter, Frédéric Monnier, and Edoardo Charbon. Single-photon synchronous detection. *IEEE Journal of Solid-State Circuits*, 44(7):1977–1989, 2009. 1.2.2

[74] Thierry Oggier, Rolf Kaufmann, Michael Lehmann, Bernhard Buttgen, Simon Neukom, Michael Richter, Matthias Schweizer, Peter Metzler, Felix Lustenberger, and Nicolas Blanc. Novel pixel architecture with inherent background suppression for 3d time-of-flight imaging. In *Electronic Imaging 2005*, pages 1–8. International Society for Optics and Photonics, 2005. 5.1

[75] Matthew O'Toole, Ramesh Raskar, and Kiriakos N. Kutulakos. Primal-dual coding to probe light transport. *ACM Transactions on Graphics*, 31(4):1–11, July 2012. ISSN 07300301. doi: 10.1145/2185520.2185535. 1.1, 1.3.1, 2.1.1, 4.1.1

[76] Matthew O'Toole, John Mather, and Kiriakos N Kutulakos. 3d shape and indirect appearance by structured light transport. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3246–3253. IEEE, 2014. 1.1, 1.3.1, 1.3.5, 4.1.1, 4.2, 4.3.1, 5.2

[77] Matthew O'Toole, Supreeth Achar, Srinivasa G Narasimhan, and Kiriakos N Kutulakos. Homogeneous codes for energy-efficient illumination and imaging. *ACM Transactions on Graphics (TOG)*, 34(4):35, 2015. 1.4, 4.2, 5.1

[78] Jordi Pagès, Joaquim Salvi, Christophe Collewet, and Josep Forest. Optimised De Bruijn patterns for one-shot shape acquisition. *Image and Vision Computing*, 23(8):707–720, August 2005. ISSN 02628856. doi: 10.1016/j.imavis.2005.05.007. 1.3.5

[79] Andrew Payne, Andy Daniel, Anik Mehta, Barry Thompson, Cyrus S Bamji, Dane Snow, Hideaki Oshima, Larry Prather, Mike Fenton, Lou Kordus, et al. A 512× 424 cmos 3d time-of-flight image sensor with multi-frequency photo-demodulation up to 130mhz and 2gs/s adc. In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2014 IEEE International*, pages 134–135. IEEE, 2014. 1.2.2

[80] A. P. Pentland. A new sense for depth of field. *IEEE Trans. Pattern Anal. Mach. Intell.*, 9(4):523–531, April 1987. ISSN 0162-8828. doi: 10.1109/TPAMI.1987.4767940. URL `http://dx.doi.org/10.1109/TPAMI.1987.4767940`. 1, 1.3.2

[81] Georg Petschnigg, Richard Szeliski, Maneesh Agrawala, Michael Cohen, Hugues Hoppe, and Kentaro Toyama. Digital photography with flash and no-flash image pairs. *ACM transactions on graphics (TOG)*, 23(3):664–672, 2004. 1.1, 1.3.5

[82] Ramesh Raskar, Kar-Han Tan, Rogerio Feris, Jingyi Yu, and Matthew Turk. Non-photorealistic camera: depth edge detection and stylized rendering using multi-flash imaging. In *ACM transactions on graphics (TOG)*, volume 23, pages 679–688. ACM, 2004. 1.1

[83] Ramesh Raskar, Amit Agarwal, and Jack Tumblin. Coded Exposure Photography : Motion Deblurring using Fluttered Shutter. *ACM Transactions on Graphics*, 25(3):795–804, 2006. 2.5

[84] D. Reddy, A. Veeraraghavan, and R. Chellappa. P2c2: Programmable pixel compressive camera for high speed imaging. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 329–336, June 2011. doi: 10.1109/CVPR.2011.5995542. 4.3.1

[85] Dikpal Reddy, Ravi Ramamoorthi, and Brian Curless. Frequency-Space Decomposition and Acquisition of Light Transport under Spatially Varying Illumination. *European Conference on Computer Vision*, 2012. 1, 1.1

[86] Navid Sarhangnejad, Hyunjoong Lee, Nikola Katic, Matthew O'Toole, Kiriakos Kutulakos, and Roman Genov. Primal-dual-coding cmos image sensor architecture. *International Image Sensor Workshop*, 2017. 7.1, 7.1

[87] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vision*, 47(1-3):7–42, April 2002. ISSN 0920-5691. doi: 10.1023/A:1014573219977. URL `http://dx.doi.org/10.1023/A:1014573219977`. 1

[88] YY Schechner and Nahum Kiryati. Depth from defocus vs. stereo: How different really are they? *International Journal of Computer Vision*, 39(2):141–162, 2000. 1.3.2, 3.1.1

[89] Horst Schreiber and John H Bruning. Phase shifting interferometry. *Optical Shop Testing, Third Edition*, pages 547–666, 2006. 1.1

[90] Rudolf Schwarte, Zhanping Xu, Horst-Guenther Heinol, Joachim Olk, Ruediger Klein,

Bernd Buxbaum, Helmut Fischer, and Juergen Schulte. New electro-optical mixing and correlating sensor: facilities and applications of the photonic mixer device (pmd), 1997. URL http://dx.doi.org/10.1117/12.287751. 6

[91] S.M. Seitz, Y. Matsushita, and K.N. Kutulakos. A theory of inverse light transport. *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, pages 1440–1447 Vol. 2, 2005. doi: 10.1109/ICCV.2005.25. 1.3.1

[92] Pradeep Sen, Billy Chen, Gaurav Garg, Stephen R Marschner, Mark Horowitz, Marc Levoy, and Hendrik Lensch. Dual photography. In *ACM Transactions on Graphics (TOG)*, volume 24, pages 745–755. ACM, 2005. 1.1

[93] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. Real-time human pose recognition in parts from single depth images. *Commun. ACM*, 56(1):116–124, January 2013. ISSN 0001-0782. doi: 10.1145/2398356.2398381. URL http://doi.acm.org/10.1145/2398356.2398381. 5.1

[94] V Srinivasan, H C Liu, and Maurice Halioua. Automated phase-measuring profilometry : approach. *Applied Optics*, 24(2):185–188, 1985. 1.2.3, 1.3.5

[95] AB Stanbridge and DJ Ewins. Modal testing using a scanning laser doppler vibrometer. *Mechanical systems and signal processing*, 13(2):255–270, 1999. 1.1

[96] Frank Steinbrücker and Daniel Cremers. Large Displacement Optical Flow Computation without Warping. *ICCV*, 2009. 2.1.1, 2.3.2

[97] Shuochen Su, Felix Heide, Robin Swanson, Jonathan Klein, Clara Callenberg, Matthias Hullin, and Wolfgang Heidrich. Material classification using raw time-of-flight measurements. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 5.2, 6, 7.1

[98] Bo Sun, Ravi Ramamoorthi, Srinivasa G Narasimhan, and Shree K Nayar. A practical analytic single scattering model for real time rendering. *ACM Transactions on Graphics (TOG)*, 24(3):1040–1049, 2005. 7.1

[99] Tracey T. Sutton. Trophic ecology of the deep-sea fish malacosteus niger (pisces: Stomiidae): An enigmatic feeding ecology to facilitate a unique visual system? *Deep Sea Research Part I: Oceanographic Research Papers*, 52(11):2065 – 2076, 2005. ISSN 0967-0637. doi: https://doi.org/10.1016/j.dsr.2005.06.011. URL http://www.sciencedirect.com/science/article/pii/S0967063705001652. 1

[100] Ryuichi Tadano, Adithya Kumar Pediredla, and Ashok Veeraraghavan. Depth selective camera: A direct, on-chip, programmable technique for depth selectivity in photography. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3595–3603, 2015. 5.2

[101] Yuichi Taguchi, Amit Agrawal, and Oncel Tuzel. Motion-Aware Structured Light Using Spatio-Temporal Decodable Patterns. *European Conference on Computer Vision*, 2012. 1.3.5

[102] Dharmpal Takhar, Jason N. Laska, Michael B. Wakin, Marco F. Duarte, Dror Baron, Shri-

ram Sarvotham, Kevin F. Kelly, and Richard G. Baraniuk. A new compressive imaging camera architecture using optical-domain compression. In *Proceedings of Computational Imaging IV at SPIE Electronic Imaging*, pages 43–52, San Jose, CA, Jan. 2006. 4.3.1

[103] Michael Tao, Sunil Hadap, Jitendra Malik, and Ravi Ramamoorthi. Depth from Combining Defocus and Correspondence Using Light-Field Cameras. *International Conference on Computer Vision*, 2013. 3.1.1

[104] Hui Tian. *Noise analysis in CMOS image sensors*. PhD thesis, Stanford University, 2000. 1.3.3

[105] Borom Tunwattanapong, Graham Fyffe, Paul Graham, Jay Busch, Xueming Yu, Abhijeet Ghosh, and Paul Debevec. Acquiring reflectance and shape from continuous spherical harmonic illumination. *ACM Trans. Graph.*, 32(4):109:1–109:12, July 2013. ISSN 0730-0301. doi: 10.1145/2461912.2461944. URL `http://doi.acm.org/10.1145/2461912.2461944`. 1.1

[106] Michael Weber, Michaela Mickoleit, and Jan Huisken. Light sheet microscopy. *Methods in cell biology*, 123:193–215, 2013. 1.1, 5.1

[107] Thibaut Weise, Bastian Leibe, and Luc Van Gool. Fast 3D Scanning with Automatic Motion Compensation. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2007. doi: 10.1109/CVPR.2007.383291. 1.3.5, 2.1.1

[108] Robert J. Woodham. Shape from shading. chapter Photometric Method for Determining Surface Orientation from Multiple Images, pages 513–531. MIT Press, Cambridge, MA, USA, 1989. ISBN 0-262-08183-0. URL `http://dl.acm.org/citation.cfm?id=93871.93888`. 1.1

[109] Ta Yuan and Murali Subbarao. Integration of Multiple-Baseline Color Stereo Vision with Focus and Defocus Analysis for 3-D. In *Proceedings of SPIE*, number i, pages 44–51, November 1998. doi: 10.1117/12.334349. 3.1.1

[110] Li Zhang and Shree Nayar. Projection defocus analysis for scene capture and image display. *ACM Transactions on Graphics*, 2006. 1.3.2, 3.1.1, 3.2, 3.2, 3.7

[111] Li Zhang, Brian Curless, and Steven M Seitz. Spacetime stereo: Shape recovery for dynamic scenes. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–367. IEEE, 2003. 1.3.5

[112] Li Zhang, Noah Snavely, Brian Curless, and S Seitz. Spacetime Faces: High-Resolution Capture for˜ Modeling and Animation. *SIGGRAPH*, 2004. 2.1.1