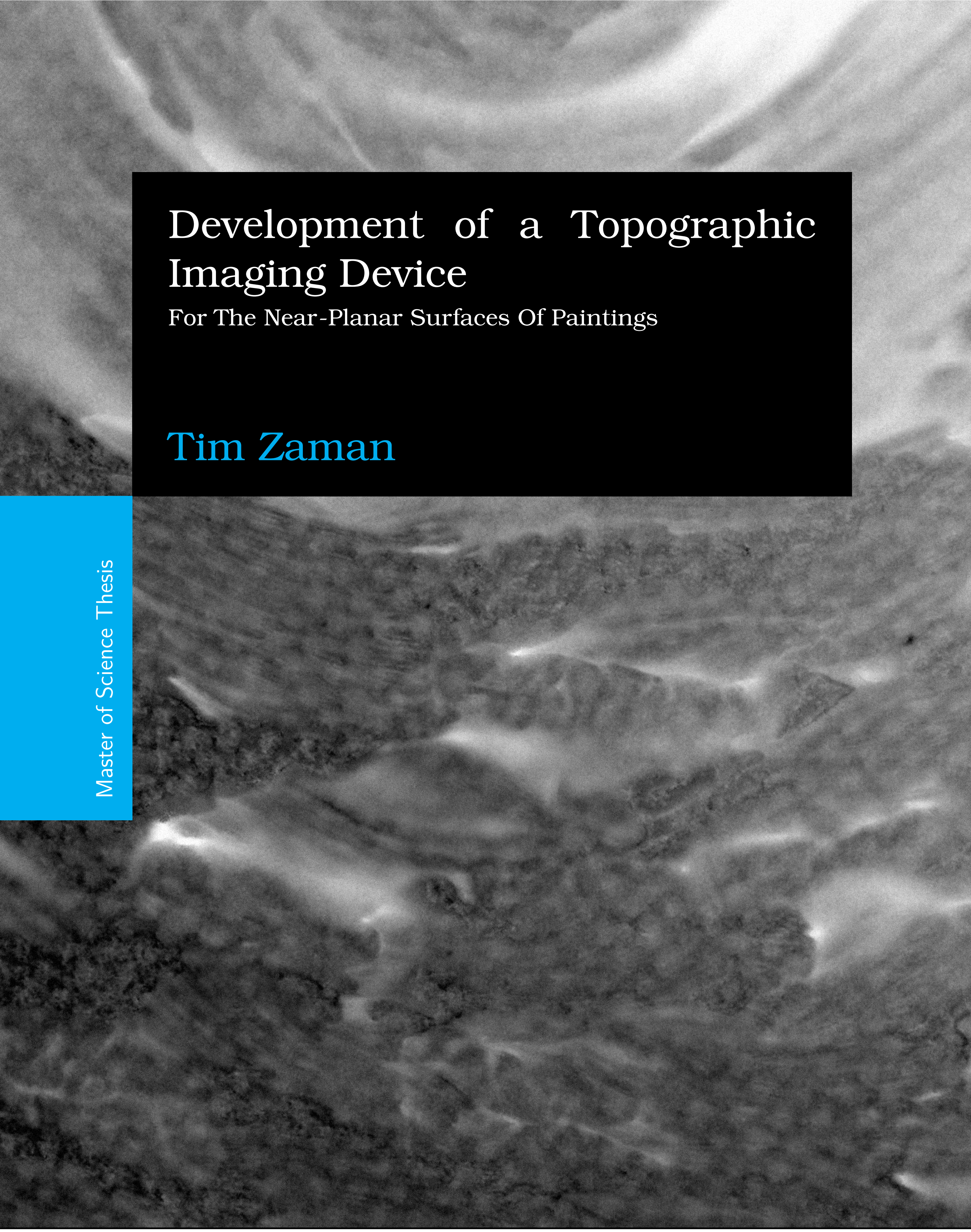


Development of a Topographic Imaging Device

For The Near-Planar Surfaces Of Paintings

Tim Zaman

Master of Science Thesis

Mechanical Engineering

Development of a Topographic Imaging Device

For The Near-Planar Surfaces Of Paintings

Master of Science Thesis

For the degree of Master of Science in Mechanical Engineering at Delft University of Technology

Tim Zaman

February 20, 2013

Faculty of Mechanical, Maritime and Materials Engineering (3mE) · Delft University of Technology

The work in this thesis was supported by Picturae BV. Their cooperation is hereby gratefully acknowledged.

Delft University of Technology
Department of
Biomechanical Engineering

The undersigned hereby certify that they have read and recommend to the Faculty of Mechanical, Maritime and Materials Engineering (3mE) for acceptance a thesis entitled

Development of a Topographic Imaging Device

by

Tim Zaman

in partial fulfillment of the requirements for the degree of

Master of Science Mechanical Engineering

Dated: February 20, 2013

Supervisor(s): Prof.dr.ir. P.P. Jonker

Dr.ir. B. Lenseigne

Reader(s): Ir. J.C. Verlinden

Prof.dr. J. Dik

Abstract

Paintings are versatile near-planar objects with material characteristics that vary widely. The fact that paint has a material presence is often overlooked, mostly because of the fact that we encounter many of these artworks through two dimensional reproductions. The capture of paintings in the third dimension is not only interesting for study, restoration and conservation, but it also facilitates making three dimensional reproductions through novel 3-D printing methods. These varying material characteristics of paintings are first investigated, after which an overview is given of the feasible imaging methods that can capture a painting's color and topography. Because no imaging method is ideally suited for this task, a hybrid solution between fringe projection and stereo imaging is proposed involving two cameras and a projector. Fringe projection is aided by sparse stereo matching to serve as an image encoder. These encoded images processed by the stereo cameras then help solve the correspondence problem in stereo matching, leading to a dense and accurate topographical map, while simultaneously capturing its color. Through high-end cameras, special lenses and filters we capture a surface area of 170 cm^2 with an in-plane effective resolution of 50 µm and a depth precision of 9.2 µm. Semi-automated positioning of the system and data stitching consequently allows for the capture of surfaces up to 1 m^2. The reproductive properties are conform the digitization guidelines for cultural heritage. The preliminary results of the capture of a painting by Rembrandt reveal that this system is indeed ideally suited as a topographic imaging device for the near-surface of paintings. This data can be used for making a lifelike reproduction in full dimension and full color.

Table of Contents

List of Figures

List of Tables

Preface

This document is a part of my Master of Science graduation thesis. The idea of doing my thesis on this subject came after a spontaneous visit to prof.dr. Joris Dik after reading an article[11] about his research in the university's newspaper. His research focuses on the development and application of new, non-destructive methods of analysis for research of works of art. The courses I attended leading up to this document were centered around my strong interest in digital imaging. During these courses and my internship[12] I have worked in an R&D position for Picturae, a company specializing in mass digitization of cultural heritage.

The subject of the thesis is therefore a synthesis of both my own experience and interest, as well as one between the Materials in Art and Archaeology section and the Vision Based Robotics section (under the supervision of prof.dr.ir. Pieter Jonker) of the Delft University of Technology.

Joris Dik expressed the interest in having the ability of accurately and simultaneously scanning the color and topography of paintings, for which no proven out-of-the box solutions exist. The topographic scans are not only useful to support current research, but also creates the ability to accurately reproduce works of art in full dimension, scale, color and resolution through new 3-D printing methods.

Apart from the previously mentioned names, I would like to express gratitude to dr.ir. Boris Lenseigne for valuable feedback in the realization of this document, my sister for grammar and spell checking and Picturae for letting us utilize their photographic equipment.

Delft, University of Technology
February 20, 2013

Tim Zaman

“Research is what I’m doing when I don’t know what I’m doing. ”

— *Wernher von Braun*

Chapter 1

Introduction

The appearance of a plane is not only affected by its color, but also by the light illuminating the object. This light makes the color apparent to us and casts shadows and reflections. Especially in (near-)planar surfaces, imperfections in its 'flatness' quickly become apparent through illumination. An (artificial) illustration of this principle is given in Figure 1-1. A shadow has the power to cast a deep black on a white surface, and a reflection can cast a highlight on a black one. These phenomena originate from irregularities in the topography of the plane.

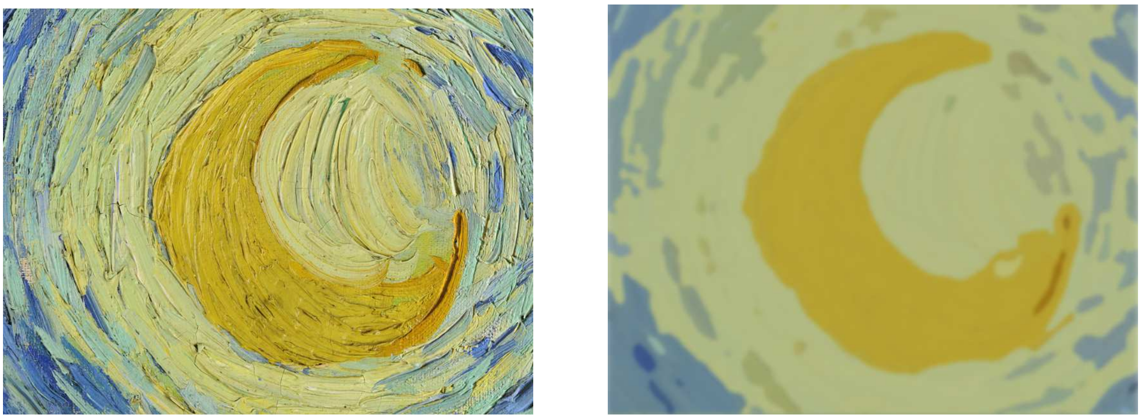
One important realization is that the difference between Figures 1-1(a) and 1-1(b) is only induced by a matter of millimetres in its topography, while the in-plane diagonal of the painting in the figure is around 15 cm, and the total painting over one metre. So although we can see the effects of this depth deviation, the depth deviation itself is hard to perceive due to the large relative scale of depth versus in-plane size.

A well known class of objects where a small depth deviation plays a crucial role in the appearance of a plane are paintings. In this document, paintings will be the planar objects of focus because their surface characteristics vary widely, and methods resulting from this overview are therefore likely to be compatible to other near-planar surfaces.

From a technical point of view, the painter induces carefully considered imperfections (paint strokes) on a planar surface (canvas) in order to create a depiction. When viewing an actual painting, the material aspects of the paint can become very apparent to the observer. These material induced effects and its dynamics are mostly imperceptible in today's two dimensional reproductions of these works.

A truly faithful reproduction of a painting can only be made when it is viewed on the same medium, like illuminated paper or canvas, and it should mimic not only its original size and color, but also its material aspect like texture and topography.

In this document, we will assess the current state of technology that can allow us to take the next step in making accurate reproductions. The focus will lay on the acquisition of the topography and color itself. We will give an overview of all viable depth perception methods, and carefully select different methods that will lead us into constructing an imaging device



(a) Painting Detail

(b) Possible Actual Color

Figure 1-1: Very small depth details can induce a very different appearance due to local very dark shadows, or bright (inter-)reflections. *Image Courtesy of the Museum of Modern Art (NY).*

ideally suited to simultaneously capture the color and topography of near-planar surfaces like paintings.

1-1 Digital Reproduction and Relevance

One of the big areas of the digital (two dimensional) reproduction of planar surfaces is cultural heritage. Most of our cultural heritage is held by museums, libraries and archives. Making this material accessible to the public is an important part of their mission, and in order to facilitate that, they digitize it and put it on the internet. Doing this has several advantages. Besides making the content accessible, it will build a dynamic database that can be easily administered. The digital files can be used as a preservation method [13], relevant for items that are slowly degrading. Digital collections often make the actual physical collection less relevant, which can be advantageous because the items do not have to be collected from inventory for inspection by individuals. This means less work for librarians, less need for the study-hall, and the items are not handled - improving their health.

The digital representation of the material that has been previously discussed is mostly nothing more than a common red, green and blue (RGB) image with some metadata. For items that are considered to be of high cultural value, great care is taken and they are frequently analyzed with many different imaging techniques and methods that can provide new insights. Methods range from fairly common infrared (IR) photography [14] to cutting edge methods like TeraHerz imaging [15] or X-ray fluorescence (XRF) spectroscopy [16]. From frequent use of raking light images and surface normal maps it is apparent that there is an interest in the capture of the surface topology, which is not often done due to the limited availability of the equipment to accommodate it. From the surface topology, these raking light images and surface normal maps can be derived as well. In general, any of these forms of digitization facilitates restauration, study, conservation and exploitation (publicity or monetary gain).

1-1-1 3-D Reproductions

The previously discussed methods are all confined to the two dimensional plane, while paintings have a three dimensional presence that is carefully considered by the artist [2]. Up until recently, this extra dimension had a limited use. Its use was limited to general topographic analysis, detecting local surface defects and other restoration purposes. Because of developments in consumer 3-D screens and printers and improvements in computation power, the relevance of 3-D scanning and publishing is increasing.

To our knowledge, no art reproductions have ever been made that are in full color and printed accurately to scale in all three dimensions before the realization of this project. New developments in high resolution 3-D printing makes such a reproduction possible, and sparked this initiative into the development of a high resolution 3-D scanner. For our purposes, we need scans of a high quality and resolution covering a large planar surface, that captures both color and topography. For these demands there are no proven systems on the market.

Recent developments in 3-D printing are increasingly allowing for high quality 3-D prints in full color. More well known and old fashioned techniques that attempt to reproduce paintings with topography have relied on reproducing the painting with real paint, of which a mold is made. This mold is then used to press the topographic shape into a 2-D print [17, 18]. The topography is here not true to the original painting, but rather a topographical reproduction of someone's painted reproduction and will therefore not be accurate.

1-1-2 Areas of 3-D Acquisition

The acquisition of environments in three dimensions spans a wide range of fields. Among those from which we can derive knowledge and experience are notably that of automotive [19], robotics [20], augmented reality [4], healthcare [21], entertainment [22], kinematics [23], surface analysis [24], reverse engineering [25], geodesy [26] and many others. Few case studies have been published that focus on the topographical digitization of paintings, but there have been many case studies on statues [1, 27, 28, 29].

1-2 Thesis Goals

This document will give an overview of the current methods in 3-D imaging that allow the accurate capture of surface topology of large planar surfaces like paintings, for the purpose of making 3-D full color, full scale reproductions. The different machine depth perception methods will be evaluated through literature for their potential to meet our design requirements. From the different methods, a viable design will be made in practice utilizing hardware and software. The resulting performance of this design is then evaluated in the context of the design constraints.

1-3 Thesis Outline

The thesis will follow a structure depicted in Figure 1-2. We will first investigate the surface of a plane, its material aspects and the optical phenomena that make up the appearance to

the observer. The human observer and its depth perception are discussed in Chapter 3.
In Chapter 4 the methods of machine depth perception are discussed and evaluated for our purposes, investigating hardware components and software solutions. The two methods of interest are stereo vision and fringe projection discussed in respectively Chapters 5 and 6. The synthesization of these two methods is then discussed in Chapter 7.
The fusion of multiple depth images is then discussed in Chapter 8, after which we propose a practical design in Chapter 9. This design is then evaluated with respect to our constraints in Chapter 10, and a conclusion is given in Chapter 11.

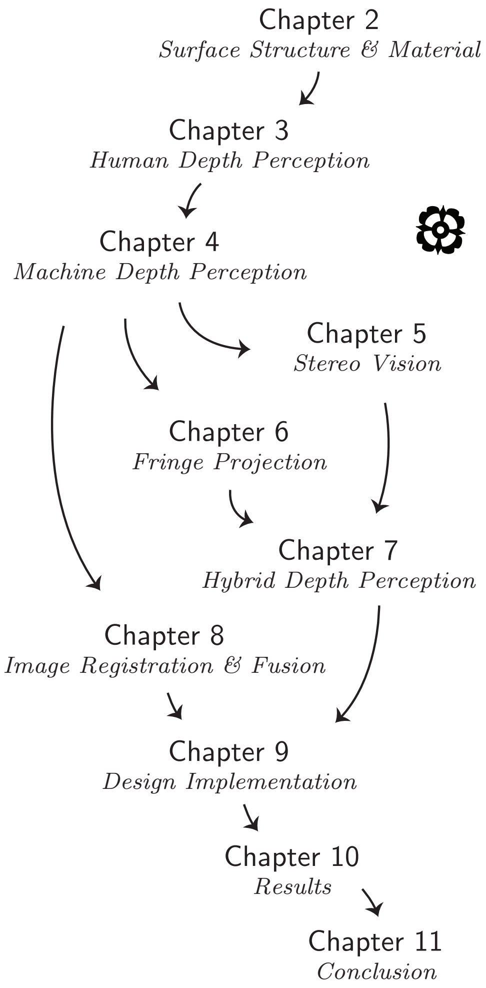


Figure 1-2: Thesis Roadmap

Chapter 2

Surface Structure and Material

As one of our goals is to reproduce the topography of a large planar surface like a painting, we investigate what a painting is made of. Browsing through any museum, you will notice that there are many differences between the surface characteristics of paintings. Size differences are obvious, but when looking consciously you will notice that some are highly reflective, while others seem matte. Their texture differs, and some are crackled or even damaged. The dynamic range (DNR) of the colors varies widely. All the complications and different aspects involved make the painting in general a complex and interesting planar surface to investigate, easily compatible with other fields of planar 3-D scanning where objects other than paintings are involved.

2-1 What Is a Painting?

A typical painting is made on a planar surface which can be anything from walls to wood to canvas. On this surface, a uniform ground layer is applied. On this layer, the final picture is painted with paint. Paint is usually made up out of two components, the binder (as medium) and the pigment (as colorant) as depicted in Figure 2-1. The thickness of the paint layer deviates widely, even within a single painting, and is often made up out of multiple layers. The cumulative thickness of the layers of paint is usually in the order of 0.1 mm to 2 mm.

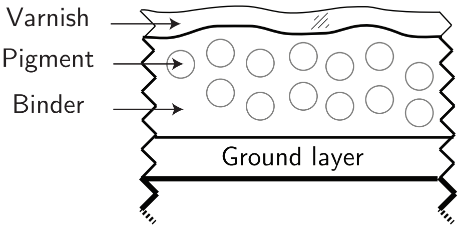


Figure 2-1: Cross-section displaying the material composition of a typical painting.

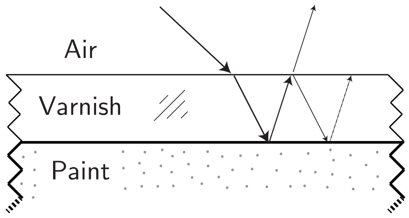


Figure 2-2: Interreflections of incident light on a layer of transparent varnish.

The pigments can be further divided into two groups, the organic and the synthetic. The pigment particles in the paint are ground to a size (roughly around the order of 0.1 µm to 10 µm) forming a uniform paste and cannot be individually identified by the human eye. A binder of oil (oil paint) is commonly used in historic paintings, although there are many alternatives like hot wax (encaustic), egg binder (tempera) or modern synthetic polymers (acrylic). Additives can also be used to change the properties of the paint like texture, opacity and reflectivity. On top of the final picture a varnish (usually around 100 µm thick [30]) is often applied, a transparent film layer. The varnish protects the painting from interacting with the environment and changes the reflective and refractive properties of the object.

2-2 Surface Reflectance

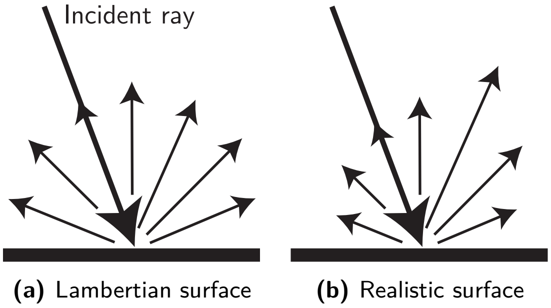
When a varnish is applied on a painting, it fills up the rougher paint surface while the surface that is exposed to the environment is smoothened due to its low viscosity. This last fact makes the rough surface's specularity homogeneous; from matte to specular, depending on the varnish used. Perceptually, the use of varnish makes the colors more vibrant. The use of varnish also comes with side effects. A reflective varnish will also immediately reflect incident light that hits its surface that can produce glares on the surface that obscure the paintings depiction. Furthermore, this light can be reflected multiple times inside of the varnish between the paint/varnish and varnish/air interface as depicted in Figure 2-2. Luckily, this phenomenon is not very dominant due to the loss of intensity with each occurrence of an interreflection.

The opposite of a highly specular mirror-like surface is ideally matte and is said to have a Lambertian reflectance. Such a surface implies that any directional light on a Lambertian surface is reflected with equal intensity in all directions as depicted in Figure 2-3. However, this does not imply that the surface appears equally bright when viewed from any angle, due to the relative inclination difference of the surface. This relation between the angle of the observer's eye and the surface is called Lambert's cosine law [31] as:

$$I_o = k_d I_p \cdot \cos(\theta).$$

In which θ is the angle of observation parallel to the surface, I_o the observed intensity, I_n the light as observed normal to the surface and k_d the Lambertian constant that is the percentage of reflected light. This relation is illustrated in Figure 2-4.

Summarizing, reflection of light from surfaces can be classified into two categories, diffuse and specular (or body and surface components). The brightness observed from a surface can be



(a) Lambertian surface **(b)** Realistic surface

Figure 2-3: We can see that a real surface has a somewhat diffuse reflection with a directional component with a higher reflectance than an ideal Lambertian surface that emits light intensity uniformly.

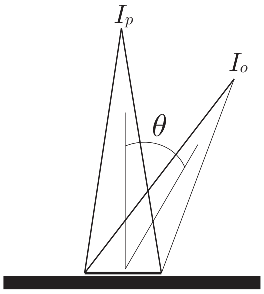
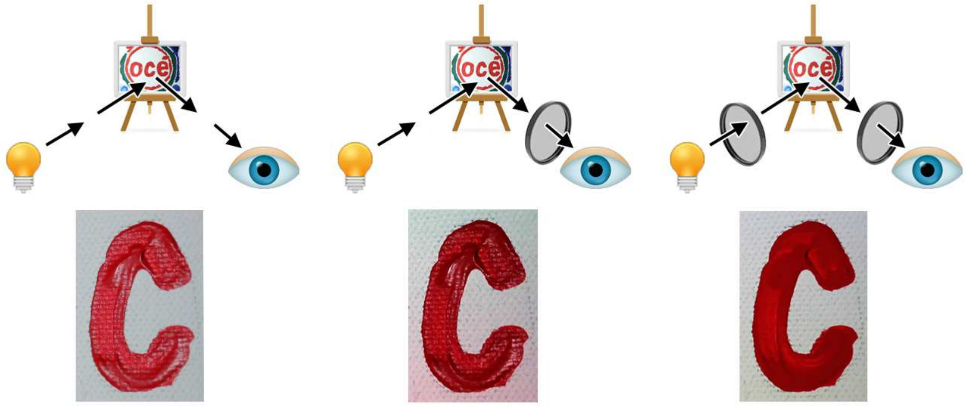


Figure 2-4: Lambert's cosine law states that there is an observed intensity (I_o) drop off with an increasing angle from the surface normal, where the intensity I_p is highest.

written as the sum of these components [32], $I = I_d + I_s$. The nature of the diffuse component is that results from light penetrating the surface in which it reflects and refracts and re-emerges. The specular component is a surface phenomenon, in which the angle of incidence equals the angle of reflection. In machine vision applications, this specular component can pose problems because the position, brightness and size of this component in the case of an image is heavily dependent on a large amount of variables, which makes it difficult to predict. We can cope with these problems in two ways: a software method in which we predict and then ignore or use this component and the laws of reflection, or we can use a hardware method in which we try to suppress this component.

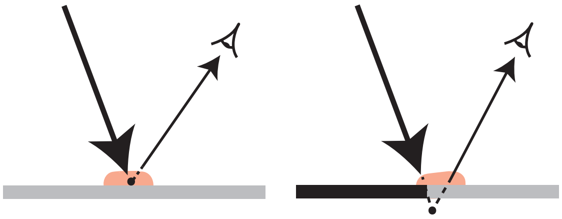
A well known method that can suppress direct reflections is the use of polarizing filters in front of the sensor of the observant. When a regular unpolarized light ray hits a nonmetallic surface, the specular part of the surface reflectance gets strongly polarized in a single direction. A polarization filter blocks light that is polarized in a certain direction, and if oriented correctly can block the light caused by the specular component very effectively [32]. A disadvantage is that the polarization filter has to be oriented correctly, and that its use will induce a brightness reduction of 4-8 times [33]. An illustration of the effect of a polarization filter is depicted in Figure 2-5. A very effective setup is called cross-polarization, where not only the surface specularities are filtered out, but specularities are prevented before the rays hit the surface that further reduce the polarization of the reflection, effectively entirely supressing the surface reflections.

Another simple and effective way to reduce specularities in structured light methods is with



(a) Regular image **(b)** Polarization **(c)** Cross-polarization

Figure 2-5: When a polarization filter is oriented correctly, specularities that appear in a regular image like (a) are mostly filtered out and we observe a response like image (b). Note that the lighting conditions in these images are identical. A cross-polarization setup is displayed in image (c).



(a) Homogeneous surface **(b)** Surface color step

Figure 2-6: Illustration of the nature of the height artifact that can occur with ray-observer triangulation on a surface that has a inhomogeneous intensity. In (a) the homogeneous case is displayed, giving an accurate surface location triangulation, while in (b), the triangulation is biased because the average position of the reflected ray has a brighter intensity on the brighter side.

the use of a linear diffuser [34]. This is only applicable in the use of structured light having a two-dimensional pattern that only varies along one dimension. The working principle behind it is that the structured light is not directly luminating in the sense of a point source as is usually the case in light sources, but it luminates the pattern from a large surface, suppressing highlights.

When capturing a surface, inhomogeneity of the surface's reflectivity can also lead to artifacts. When working with lasers or structured light, often the triangulation works with a point spread function (PSF). The single point that is triangulated is often the weighted center of this PSF. This PSF can become distorted and irregular when moving over a change of reflectivity, which results in a height effect. This effect is illustrated in Figure 2-6. With a moderate reflectivity step of 5:1, the center of gravity can shift by $\Delta x = 0.58\sigma$, with σ being the standard deviation of the PSF [35].

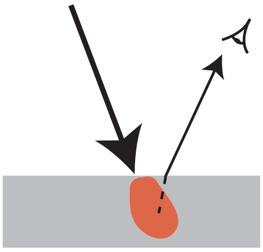


Figure 2-7: The scattered light on the opaque material forms a volume below the surface. Adopted from [1]. This leads to a biased surface location triangulation. This mostly occurs with thick and opaque materials like marble.

2-3 Subsurface Scattering

Often when a surface is diffuse and the material opaque, subsurface scattering can occur. This phenomenon was encountered in The Digital Michelangelo project, where they used a laser to scan the surface of opaque white marble [1]. This resulted in an increase of noise and a systematic bias in the derived depth, which is graphically displayed in Figure 2-7. The magnitude of the bias is proportional with the angle between that of incidence and the observer. In their marble sample, this bias was in the order of 40 µm. This bias can only be accounted for if the composition is homogeneous. In the case of paintings, it is doubtful that the composition is homogeneous. On the other hand, since paint is often darker and is much less opaque than marble, the magnitude of the bias and noise will be even lower.

2-4 Color

An integral part of the appreciation of any painting is its use of color. Technically, visible light is electromagnetic radiation with a wavelength roughly between 400 nm to 700 nm. In our case, light radiating from a painting is made up out of a combination of a large amount of photons (elementary light particles) that oscillate at different wavelengths. The light the painting reflects is dependent on the light that illuminates it. We can model the paint as either reflecting or absorbing the photons and the amount of which it does so depends on the material. The relative quantity of light that is reflected or absorbed is dependent on the wavelength, so we can model every material as having its own spectral reflectance curve as depicted in Figure 2-8. So the reason that the color of incident light on a painting is different from its reflected light, is because photons at certain wavelengths have been absorbed and usually dissipated to heat.

Remarkably, the human eye does not measure the spectral reflectance in terms of wavelengths to relative reflectance. Color is something we perceive with the cone cells in our eyes (see Section 3-1-2). In the cones, we can distinguish three receptors that are sensitive respectively to RGB. Hence, the human eye takes just three samples from the spectrum with their peaks around three different wavelengths, and constructs our color perception. We see the same trick in digital equipment like camera's and displays. These imaging devices consist out of either sensors or emitters that are sensitive to RGB, and they work well for our eyes, because our eyes have this same sensitivity. For example, in theory an RGB display can fool our

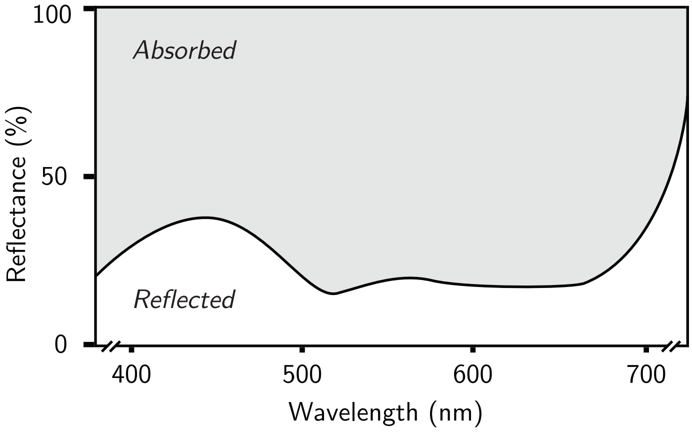


Figure 2-8: Approximate reflectance curve versus the wavelength of the pigment smalt (powdered cobalt glass). Adopted from [2].

eyes into mimicking a spectral reflectance of a material such that our eyes can not see the difference. The human eye is discussed in detail in Chapter 3.

2-4-1 Metamerism and Color Reproduction

The spectral reflectance sampling our eyes employ using the tristimulus method does have a catch. Because not all the wavelengths are separately measured, the true reflectance curve cannot be constructed. Inversely, this means that the same tristimulus perception can be constructed using different wavelengths. So two different materials with different spectral reflectance curves can lead to the perception of the same color. The catch is when we change the spectral emission from the illuminant on these two objects is changed; the different illumination leads to a different reflection. The relative spectral reflectance curve itself of the two objects does not change, but due to the different illuminant with intensities at different wavelenghts, the final absolute spectral emission from the two objects can change relative to each other. The phenomenon when colors match in one set of viewing conditions but not in others is called metamerism [36]. Metamerism is explained graphically in Figure 2-9.

The problem is that our digital displays and prints use color spaces like red, green and blue (RGB) and cyan, magenta, yellow and black (CMYK), and the exact spectral curve is not exactly replicated, only perceptually matched (metamerism is also a big problem in the retouching of paintings). Therefore, since our technology does not yet permit reproducing a certain spectral curve for our purposes, metamerism is a fundamental problem that can only be overcome by knowing the spectral emission of the illuminant and keeping it static. Even with metamerism, when the color temperature of the illuminant does not significantly change, the trisimulus approach gives a fair approximation, e.g. a red will never change into a blue.

With this knowledge, one can imagine that the lighting that is used in the exhibition of paintings is different from the one that it was painted under by the painter. The problem of metamerism in this case is overshadowed by the application of a significantly different illuminant. This museum lighting is usually done to protect paintings from a high absorption of radiation that can lead to degradation, but gives rise to the question of how important it is to the musea to display colors accurately. The fact that degradation in a reproduction

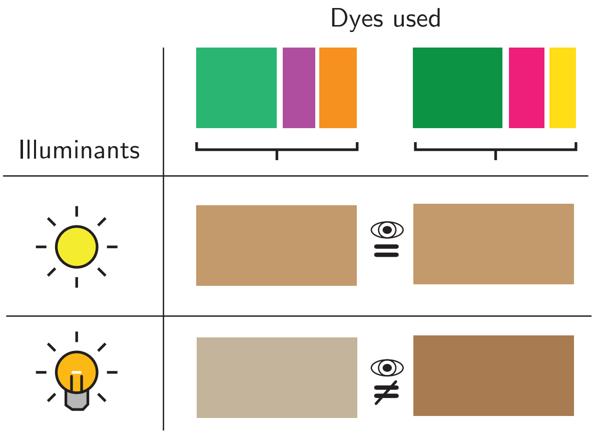


Figure 2-9: An illustration of metamerism. Colors that visually match in one viewing condition, might not match in another.

is of lesser importance, gives it the advantage that the reproduced depiction can actually be displayed more truthfully than the actual painting on display - given that the illuminant is known. Painting itself is often done in daylight, with a color temperature that can range from 5000 K for sunny daylight to 10 000 K for a heavily overcast sky. According to the Commission Internationale de l'Éclairage (CIE), a color temperature of approximately 6500 K corresponds to the standardized D65 illuminant, having the characteristics of a midday sun in Western Europe.

2-5 Reproductions

The most common method of making reproductions makes use of the ink-jet method where droplets of ink are propelled onto a surface. For high volume printing, the offset printing method is often used which is the modern variant of the printing press. The latest printers that can print the surface topology from a depth map make use of ink-jet technology, and in this document we will therefore only assume the use of this printing technology.

2-5-1 Dimensions

The physical size of the printer often puts limitations on the spatial dimensions of the reproduction. The use of rolled up surface material can extend the printable area to dozens of meters in length. Given that the printer has a very accurate positioning system for the nozzles, the size of the ink droplets and the space between them puts limitations on the spatial resolution. Often the positioning system is highly accurate, and this allows multiple layers of ink to be stacked on top of each other to allow for the construction of surface topography as displayed in Figure 2-10.

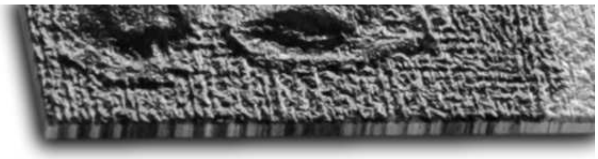


Figure 2-10: View from an angle on a stacked ink-jet print with at least 100 layers of the same depiction. From this side view you can clearly see the high accuracy and repeatability of each layer through the depiction's vertical lines on the side.

2-5-2 Color

When printers are used for reproduction, CMYK inks are often used. The input data is often encoded in the RGB color space, so these have to be converted. The printer prints using CMYK ink because it prints on a white medium, and uses the subtractive color mixing method.

In practice it is also possible to use more inks, Canon owns a printer that has 96 separate ink cartridges. This could allow for very accurate color reproduction where the spectral response can be matched. Obviously, this approach is only useful when the object's multispectral response is known, not just the RGB stimuli. Then the (spectral) characteristic of each of the inks should be known, and a model should be made for mixing these inks leading to the desired response. This degree of color reproduction is outside the scope of this document, although it is the logical next step towards the ultimate reproduction, once reproduction of the topography has been mastered.

2-5-3 Ethics of Reproduction

Our ever increasing ability to perfectly reproduce works of art raises questions to what the value is of the authentic piece of art and what the definition of authentic is. Questions like these have been posed since the industrial revolution. A renowned conclusion is that any reproduction can never reproduce the 'aura' of the painting, its 'presence in time and space', the single difference that distinguishes the original from the reproduction[37].

The possibility to reproduce the art work in 3-D could be a new step in the reproduction of art, after decades of the sole possibility of printing in 2-D. During the construction of this document, we have shown some preliminary 3-D prints to professional art conservators, with the question of what medium was used to construct 'this work', unaware that is was 3-D printed. They could accurately name what it was *not*, but the fact that is was a print did not come to mind.

Social studies on what the real impact is of this new degree of reproducibility have not been conducted. For example, one could easily imagine that strongly deteriorating and valuable works of art are put in a climate conditioned depot in order to preserve it for eternity, while the ultimate reproduction hangs in the museum. The question arises what defines the aura of an original and whether or not it can be seen in a reproduction. Especially when such a work can just as easily be bought and taken home.

Chapter 3

Human Depth Perception

Humans have the ability to perceive the world around them in three dimensions. Given the fact that the human does not employ direct 3-D sensors, it is a remarkable feat. It is relevant to investigate the methods that the human employs to achieve this, as most — if not all — of these tricks could be implemented in 3-D machine vision applications as well. This chapter gives an overview of the relevant sensors the human has that it can use in sensing depth perception, followed by how they can be implemented to actually perceive it.

3-1 Human Sensory Organs

In order to achieve depth perception, the human has multiple sensors and processing mechanisms at its disposal. The eye as a sensor is the most obvious, while its control and vestibular functions and brain processing are less apparent or even subconscious. Before depth is perceived, it first has to be sensed. In order to understand the nature of these cues, it is relevant to provide a brief summary of the most relevant sensors.

3-1-1 The Vestibular System

In order to better relate the position and movement of the human body, head and eyes to the surroundings, we need a good sense of spatial orientation. The vestibular system (part of the inner ear, one sensor system for each ear) senses rotational movements and linear accelerations. For our visual system, they have several functions. The spatial orientation helps us determine our position with respect to the earth, which can be important in some depth cues like height of the visual field (see Section 3-2). There are also direct neural feedback loops that account for the positioning of our eyes, enabling us to fixate our eyes on an object even if we change the position of our eyes.

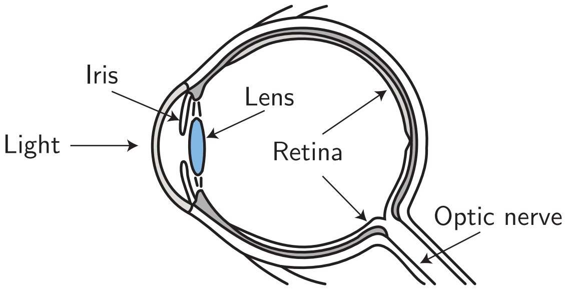


Figure 3-1: A cross-section of the human eye.

3-1-2 The Eye

The organ that enables human vision are the eyes. The eye is sensitive to electromagnetic radiation with a wavelength roughly between 400 nm to 700 nm. In Figure 3-2 we can see the gamut of colors the human eye can distinguish inside a chromaticity diagram. Around the edges of this locus are the pure wavelengths we can observe, and inside the horseshoe shape are the combinations of colors we observe through the mixture of colors. Observe that the lower edge between blue and red contains the color purple. For the color purple, there is no corresponding single pure wavelength. Our eyes perceive purple as the mixture of the outer edges of the spectrum, blue and red. In the same figure, a triangular RGB gamut is displayed. This is how a computer encodes this color data to the RGB color space. We can see that for this particular RGB profile (eciRGBv2), some color data falls outside the digital encoding regions. This is because of the trade-off in digital storage where a larger range makes the resolution (or steps) within that range smaller.

The sensitive cells are the rods and cones, responsible for respectively the lower and higher brightness sensitivity. These cells reside in the retina, which covers the inner surface of the eye as depicted in Figure 3-1. The eye is very adaptive to different brightness levels because of both neural adaptation in the retina and dilation of the pupil through contraction of the iris. The total field of view is large, for both eyes more than 180° horizontally. Only a small part of the retina contains a high density of photo-receptive cones, through which we can achieve a high local resolution. In a machine vision analogy, we can achieve a high resolution image of our environment only by scanning the environment by rotating or translating the observer. This scanning movement is done either in a smooth pursuit fashion or saccadic movements, intermittently moving and stopping. In a machine analogy, this could be seen as frame grabbing and the fusion of these images as stitching. These frames help the brain map the environment in a texturized 3-D dimensional map. Other than digital data storage, the data in our brains is highly subjective. Moreover, small saccadic movements (micro-saccades) enable the eyes to resolve features smaller than the individual size of the sensors in our retina. In machine vision, this is called super-resolution.

3-1-3 The Brain

Sensory information is processed by the Central Nervous System (CNS). In the case of the eyes, information is processed in cerebellum and the visual cortex, part of the cerebral cortex, which resides in the back side of the brain. The visual cortex can be subdivided into different

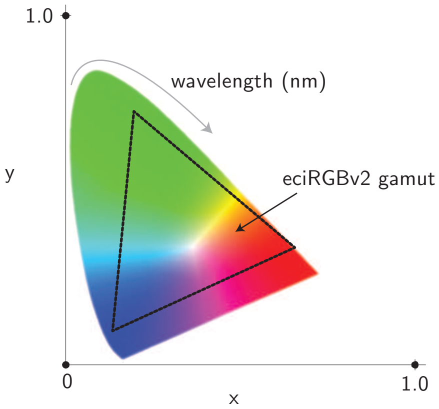


Figure 3-2: CIE Chromaticity diagram showing the gamut of colors visible to the human eye, and inside of that the triangle visible when these colors are encoded by a camera to an RGB colorspace, in this case the eciRGBv2 profile.

cortices through which the signals are subsequently processed. In a machine vision analogy, each of these sub-cortices has a higher image processing function ranging from early and fast local contrast encoding to sophisticated Object Recognition Memory (ORM).

There are two signal pathways: the dorsal stream and the ventral paths. The dorsal or 'where/how' path is responsible for object localization, motion and propriocepsis. The ventral or 'what' path is responsible for object recognition and matching and the ORM.

From the available literature on the subdivision of different visual processing functions and areas it becomes apparent that depth perception does not only arise out of the obvious binocular matching, but human depth perception also turns out to be highly redundant and to employ the fusion of sensory information.

3-2 Depth Cues

From the information the human acquires from its sensors and knowledge, the human can eventually perceive depth. There is a wide variety of cues available for the human to accommodate this, cues which the human often combines in order to get a more reliable depth estimate. These cues can be ordered in many different ways, like ordering it by sensory organs as illustrated in Figure 3-3.

Independent cues generally have a specific window of perception in which they work, and within that window they have just-discriminable thresholds in depth, as depicted in Figure 3-4. In this graph, the distance from the observer is segmented into three depth windows: the personal space, the action space and the vista space. Finally, for all cues it should be noted that depth can only be perceived for parts that are projected into the eyes of the observer. A full 3-D map or model of the environment can only be made by combining the information from the cues (projected depth or 2.5-D) and moving the observer around those objects in order perceive an all-embracing 3-D model.

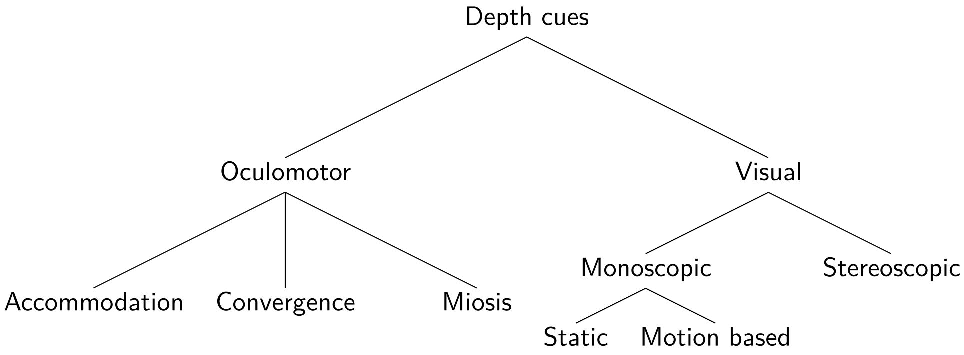


Figure 3-3: Depth cues humans use in their depth perception. Adopted from [3].

3-2-1 Occlusion

Occlusion is the most basic form of depth perception, and the earliest depth cue an infant learns [5]. It occurs when one objects obscures another, and thus does not give direct absolute or relative depth information, but rather reveals the depth order.

3-2-2 Relative Size and Density

If the size of an object in view is known, or if there are multiple objects that the observer identifies as equal, their sizes can be related to give a cue about their depth. The relative density cue arises when multiple identical objects are placed in view with regular intervals; then density scales with depth.

3-2-3 Height in Visual Field

When the observer is surrounded by a relatively flat environment, remote objects appear higher in the plane when the surface is below the level of the eye.

3-2-4 Aerial Perspective

Generally, every medium that we view our scene through contains particles that slightly obscure our view. The amount of this slight obstruction scales with depth, and therefore remote objects can appear with decreased contrast relative to objects in the foreground.

3-2-5 Motion Perspective

When objects move in relation to the observer, the projection of this movement moves over the retina in a speed that is related to the distance to the object. When this is caused by observer movement in a stationary environment, this is called motion parallax, while the motion of the objects in the field is called motion perspective [38]. Motion parallax is important for the depth perception of animals that have little visual overlap between both eyes, limiting their stereopsis.

3-2-6 Binocular Disparity, Stereopsis and Diplopia

Two dissimilar images from each eye can be combined and the relative positions of each feature in both images can be estimated. On the location where the eyes have converged on, there is no disparity; this position is in the center of the retina on each eye. Because the eyes observe from different directions, features in the scene that are lateral to the placement of the eye will have a different relative location on the projection. Objects that lie further from the horopter (zero-disparity plane) have a larger disparity and can be both negative and positive. There is also a distinction between subtle disparities and large disparities. Small disparities give the impression of solid space, and is called stereopsis. When the disparity is large, it induces double vision, called diplopia [39].

3-2-7 Convergence and Accommodation

The depth on which the eye focuses is related to the shape of the lens. This focusing by changing the shape of the lens is called accommodation. The amount of accomodation gives a depth cue. Both eyes usually focus on and point to the same point in space. Since the eyes have a fixed disparity, through the angle at which the eyes converge, the depth can be triangulated. In Figure 3-4 we can see that the just-discriminable depth threshold of this cue is very coarse, and we can see that the binocular disparity reveals a better depth discrimination and is therefore usually dominant.

3-2-8 Texture Gradients

There are at least three methods by which texture gradients can help in perceiving depth; size, density and compression. The gradients of texture size and density are similar to respectively relative size and relative density, covered in Section 3-2-2. Their only difference is that the thing that changes is the texture instead of the object itself. From texture compression (also called texture distortion) one could also derive depth information, although this information is better used in observing the shape and curvature of the object.

3-2-9 Kinetic Depth and Gravity

From the movements of an object alone, it is possible to extract depth information. This is called kinetic depth, although it reveals more information on the shape of the object than its depth. Depth from gravity should be possible because through natural gravity, absolute size and distance could be identified, although the contribution of this cue seems negligible [40]. Alternatively, from the combination of inertia and acceleration depth could also be extracted, i.e. an airplane that seems to accelerate quickly through the sky is probably a scale model.

3-2-10 Linear Perspective

A well known and strong cue for estimating depth is linear perspective, which is where we trace receding convergent perspective ‘lines’ and from it derive three dimensional information. This cue is actually not an independent cue [5], but rather a method describing depth perception

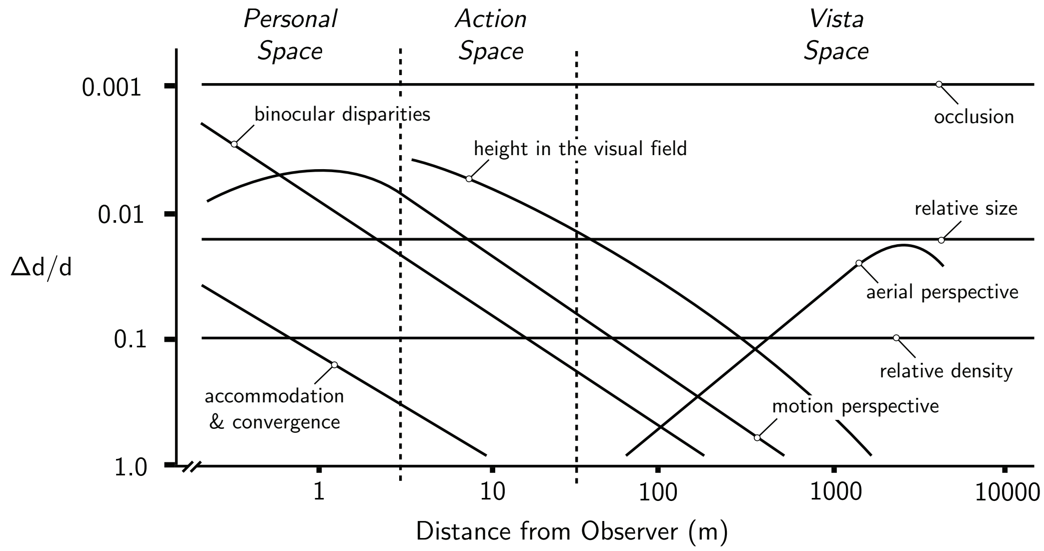


Figure 3-4: Just-discriminable depth thresholds as a function of the log (average) distance from the observer. Adopted from [4] and [5]. This graph is meant as indication, for each curve is very dependent on a large amount of variables.

that uses different independent cues like texture gradients, relative size, relative density and occlusions.

3-2-11 More Cues

Although dominant, depth cues in humans are not necessarily visual. Propriocepsis and touch sensitivity can be combined in order to get a good three dimensional perception of the environment, a cue we often use when our vision is impaired. Like other animal species such as bats, the human is also capable of echolocation [41], although its window of perception is small and inaccurate. One could also derive a depth cue revealing relative macro-scale surface topography from the abundant touch receptive sensors in the finger tips.

3-2-12 Conflicting Information

Because of the amount of depth cues available to the human, conflicting or different information from these individual cues is inevitable. This information can either strengthen a belief when their likelihood functions resemble, or they can confuse (decrease likelihood) when they are too different [42]. It is relevant to see how the human copes with this because hybrid solutions for our intended application are also feasible.

When there is information from different cues, generally the cue that has the smallest just-discriminable depth threshold is dominant [5] as displayed in Figure 3-4. This seems not to be the case in all situations. Many different mixture models and probabilistic structures have been proposed. When two cues are conflicting such that their likelihood function overlap, the cues are mixed into a new combined likelihood function [42]. The cue-combinations strategy for some cues will be optimal (in the definition of Bayesian statistics) [43], and combined cues will then actually lead to a more accurate estimation. In the case where the likelihood functions have very little or no overlap (strongly conflicting cues), a winner-takes-all function will mimic that of the cue with the least uncertainty, albeit with a lower magnitude of likelihood.

More generally, the integration of multiple depth cues can interact in different ways that are not mutually exclusive [44]:

- **Accumulation** is when the likelihoods of the cues can be combined in ways such as probability summation.
- **Cooperation** of different cues means there is a synergy between different cues, which is often the case in noisy information, for example with nonlinear integration of multiple cues.
- **Disambiguation** can occur when one cue disambiguates a likelihood of another.
- **Veto** in the case of a single unequivocal likelihood unchallenged by other cues.

Chapter 4

Machine Depth Perception

Through a combination and integration of sensors and software we have the ability to make a digital depth representation of the environment. Like the human, we have many different sensors and combinations of physical principles that we can use to achieve this. This chapter gives an overview of how digital depth information can be stored and displayed, followed by the different hardware that is often utilized in this field. We proceed by investigating a range of methods that integrates these hardware into an architecture that allows us to perceive depth. Finally we take a look at existing solutions that have been used in practice for purposes similar to ours.

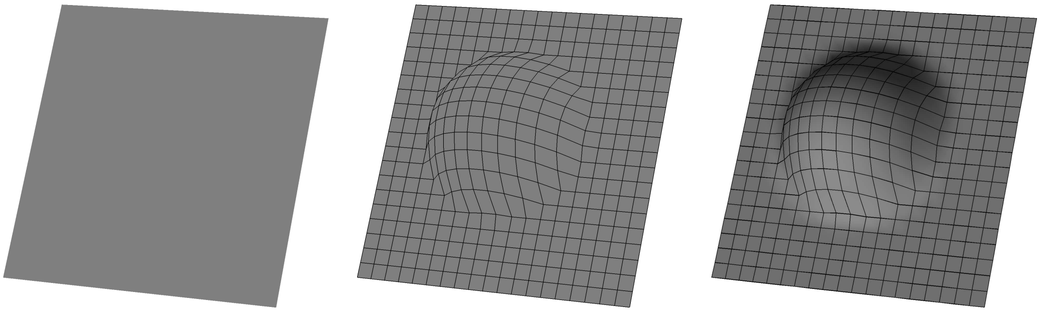
4-1 Depth Information and Representation

For two dimensional projections like images, an (x, y) mesh with a fixed interval is often chosen. Within each point (or pixel) the signal information is stored. This standard architecture is well established, memory efficient and easy.

4-1-1 Depth Information

Storing 3-D information is less straightforward, especially when the data comes from sensors. The measured points generally do not easily fit within a standard mesh, and therefore their location in each dimension (x, y, z) is stored. If more information is known, like their color, it is appended to this point. In digital 3-D graphics, these points could be displayed as a point-cloud, where each point is represented by a 'dot' in space. However, this does not give a continuously accurate representation of a real 3-D structure. To give a more accurate and continuous approximation, vertices are used. A vertex is a special point that describes edges or corners of geometric shapes. Through a large amount of these geometric polygons, an accurate approximation of any real 3-D object can be made.

In the case of the topography of a plane, we do work with 3-D data, but we assume that for each point on the (x, y) plane, there is only one depth variable z. This means that we could



(a) No depth cues (b) Imposed grid (c) Artificial lighting

Figure 4-1: Different methods of displaying the same 3-D entity, a half of a sphere on a flat surface with uniform surface characteristics. Note that all depictions are ambiguous.

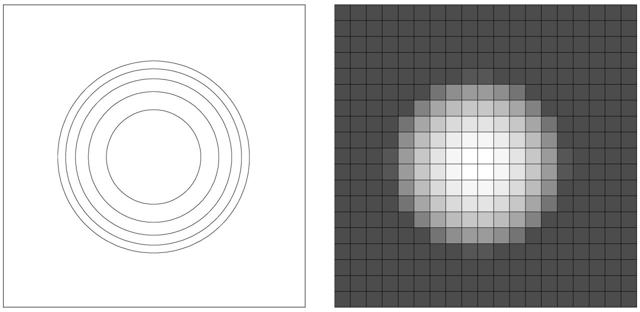
still make a two dimensional mesh with regular intervals similar to that often used in 2-D projections. For the signal value on each point on this mesh, we could use the depth value z. This is convenient, as other information like color about this point could be appended to this point as well. Such a 3-D representation on a fixed 2-D mesh is sometimes called 2.5-D [45]. Working with such representations can be very convenient. For example, computing the vertices is very straightforward, which makes 3-D rendering easy, as displayed in Figure 4-1(c).

4-1-2 Depth Representation

Three dimensional surfaces can be displayed using different methods. We investigate how this can be done in a way that the human intuitively perceives this representation objectively and unambiguously. To give a brief overview, we consider a 2.5-D planar surface with a spherical bulge as displayed in Figure 4-1.

In the most basic case, when the material is uniform and Lambertian without environmental influences, it is practically impossible to perceive depth from a surface like that in Figure 4-1(a), as there are no depth cues, apart from the outer profile of the structure. If we make the fixed mesh interval visible and superimpose it on the surface, the surface structure already becomes apparent. If we then introduce directional lighting in the environment, we can see the formation of shadows, which gives a more realistic depiction of the surface. Note that without more cues, this surface is still ambiguous (isometric illusion) and is hard to assess objectively.

Let us now consider a top-view of the surface, so that the mesh and its elements are square. This depiction is less intuitive, but more objective and unambiguous. A method that is often used in geography is the depiction of a contour map (or isolines), where curves trace constant heights. A contour map of our sample is displayed in Figure 4-2(a). A method that is frequently used in digital imaging and vision is a depth map. In the depth map in Figure 4-2(b) locations that are closer to the observer are depicted brighter, although the color map is not necessarily in gray-scale as displayed here. Although this depiction might be less intuitive than rendering methods with artificial lighting, it is preferred due to its objectiveness and unambiguity.



(a) Contour map **(b)** Depth map

Figure 4-2: Different 2-D representations of the topography (third dimension) of the planar surface displayed in Figure 4-1.

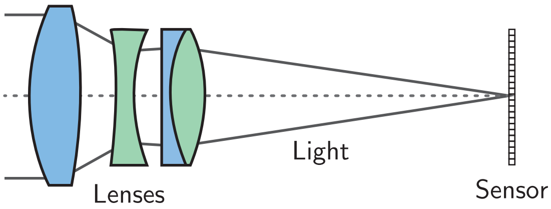


Figure 4-3: A simplified cross-section of a typical digital camera.

4-2 Hardware

The hardware sensors and components are an integral part of the machine depth perception solution. It is therefore important to understand what they do and how they work. For example, concepts like 'image capture' with a camera or 'projecting' with a projector are non trivial tasks which define the potential and constraints that the eventual system will inherit.

4-2-1 Cameras

The machine analogy of the human eye is the camera. The digital camera can be used in a variety of different setups that facilitate depth perception. In Section 3-1-2 we have seen that in essence the eye can be divided into two important parts; the lens and the sensor (the retina). A digital camera is no different, as depicted in Figure 4-3. A system of lenses focuses the incident light rays on the sensitive sensor plane, which converts these into electronic signals that are digitized and stored.

Sensor

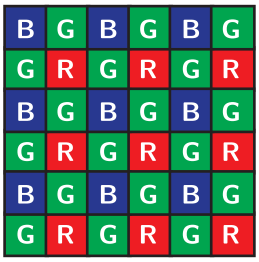
There are two main different technologies in normal digital image capture, the Charge-Coupled Device (CCD) and the complementary metal oxide semiconductor (CMOS) sensor. They share a purpose, but are different in architecture and production. In this case, the most relevant difference between these types of sensor is that a CMOS sensor with an electronic (rolling) shutter can cause artifacts due to the speed of the capture of a single frame when there is a relative difference in motion between the observer and the scene. This issue is similar to projectors that make use of a colorwheel as will be discussed in Section 4-2-2. The analog signals from the photosensitive elements are digitized through an A/D conversion, often with an 8-bit significance (256 possible values). This amount is very few if you compare this to the amount of pixels a camera usually captures in one frame (millions). Capturing color information will yield $(2^8)^3 = 2^{24} = 16,7$ million possible values for each pixel. This might seem a lot, but the brightness information we get is still effectively just 8-bit, which is enough for a human to perceive a smooth gradient for example, but can be few for machine vision.

The spectral sensitivity of the CCD and CMOS sensor is also different, as it is different from that of the human eye. The sensors have a very wide spectral sensitivity that easily encompasses that of the human eye. This means that ultraviolet (UV) and infrared (IR) radiation is also received, and will be visible in the final images. The result of this is that the environment might be perceived with a different brightness or color due to the wider spectral sensitivity, which can be overcome by using cut-filters in front of the sensor. It is important to realize that this spectral sensitivity is monochrome in nature. Even with filters, captured gray-scale images might not give an accurate gray-scale response if they are not exactly spectrally similar to the way a human interprets the brightness of different hues.

The capture of color can be achieved achieved in the same tristimulus approach as the human captures color. In the same method as the UV and IR wavelengths are filtered, one could also use filters to resemble the spectral response of the three cones that are responsible for color vision of the human eye. This is usually achieved by either a trichroic prism or a mosaiced filter array. The trichroic prism splits the light into the three (RGB) different spectral responses, which fall on three different CCD sensors (3CCD). The other method involves just one sensor on which a mosaiced filter is placed within which the mosaic is made up out of alternating red, green and blue filters (Bayer's arrangement) as depicted in Figure 4-4. A major difference in the eventual image is that with a well aligned 3CCD setup, the red, green and blue images can be placed directly on top of each other and the spatial sampling of each pixel of the projection is aligned. In the case of the mosaiced filter, the spatial sampling of the primaries is inherently misaligned, which can be subdued through interpolation. This in turn can lead to moiré patterns when the spatial frequency of the scene projected on the sensor is finer than the spacing between the pixels in the sensor [46]. The moiré effect, in turn, can be overcome by using an aliasing filter, which blurs the image slightly.

The moisaiced filter approach is by far the most common in digital color cameras. As stated before, the spatial sampling of the three color components is misaligned, and is therefore interpolated afterwards to give one RGB value for each single pixel. The method in which this is done actually preserves the amount of pixels. In other words, each pixel will have three corresponding values of RGB, while only one is sampled due to the filter. Practically this means that when a camera is stated to sample for example '10 megapixels (of RGB data)', it will actually sample 5 million green pixels, and 2.5 million green and 2.5 million blue pixels. This in turn is interpolated to yield 10 megapixels of red, green and blue data. In other words, 2 out of 3 color values are always interpolated.

In all cases, the spectral responses of the cameras sensors will in practice be different than that of the human, and therefore colors will have to be calibrated. This can be done within the processing of the images in the camera, but cannot account for all situations like differences in white balance or exposure.



B	G	B	G	B	G
G	R	G	R	G	R
B	G	B	G	B	G
G	R	G	R	G	R
B	G	B	G	B	G
G	R	G	R	G	R

Figure 4-4: A Bayer filter of which each patch within the mosaic should be directly on top of one photosensitive element. This specific arrangement is called BGGR. Note there are twice as many green filters as there are blue an red, this is because the human eye (through evolution) is relatively more sensitive to green.

The sensor system set constraints on the physical pixel size and amount of pixels (and active projection area) as well as the dynamic range, capture rate, processing speed, noise (ISO setting) and spectral sensitivity. It also influences the depth of field (DOF): with increasing sensor size the depth of field will decrease for a certain aperture, provided that the frame and distance of the projection is the same. Larger sensors would require the camera to get either closer to the subject or use a longer focal length to fill the same frame, leading to a decrease in depth of field. In the case of an electronic shutter, it defines how and when the image is captured in time.

Optical System

The optical system of the camera is responsible for how and when the light rays are guided onto the surface of the sensor. This optical system can usually be subdivided into a lens (system), a diaphragm and a (mechanical) shutter. The lens system guides the light rays to the sensor and can contain a large amount of lenses of different shapes and sizes. The diaphragm is the part that contains the aperture. The aperture is the hole through which the light rays travel and the size of this opening determines the degree of collimation of these rays. The shutter controls the time that the sensor is exposed through light rays, which can be a mechanical solution or an electronic one.

In its most minimal form, the optical system can be distilled to a tiny hole and a shutter, with no lens and no separate diaphragm. This system is called the pinhole camera model, and all light that falls on the center goes through this pinhole point in the optical center. This means we can describe the transformation of our world coordinate system to our projected coordinate system on the sensor with a geometric (perspective) model as graphically depicted in Figure 4-5. A pinhole model is therefore a simplified camera model with a very small aperture size (a high dimensionless 'f' number) without regarding (radial) lens distortion, chromatic aberration and vignetting - features that are present in practice. These distortions can be accounted for through a camera resectioning procedure as described in Section 5-2.

Apart of defining the distortions, the lens system sets the parameters as the focal length and aperture, which with the sensor constrain the field of view, depth of field and magnification. A mechanical shutter exhibits the same limitations as an electronic one. A practical camera

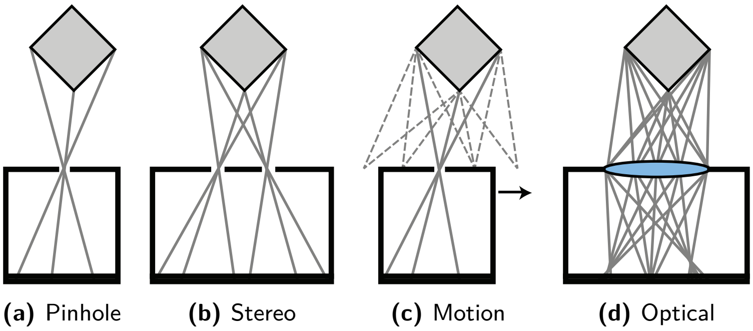


Figure 4-5: Example of projected geometry of a square object through a (a) single pinhole, (b) stereo pinhole from different viewpoints and (c) motion parallax from adjacent viewpoints. In (d) an optical system with a lens is depicted that gathers light from a continuum of viewpoints. Note that not all points are in focus as the multiple ray projections accumulate on the sensor. Adopted from [6].

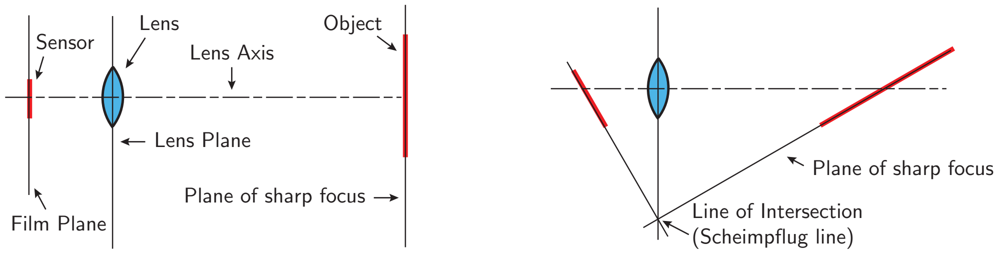
system is often dynamic in which many systems are (automatically) tuned and work together to get a good image. It is important to realize that this tuning process can influence the cameras characteristics and parameters, and would be in need of constant recalibration if applied to position-sensitive imaging algorithmic.

If the sensor plane is not perpendicular to the surface, it is important that the depth of field will not only play a role in the out-of-plane direction, but also in the height or width of the surface itself. Such setups are not uncommon when using structured light or multiple cameras. This problem can be solved if we adhere to the Scheimpflug condition [47]; then the projection of the surface will be sharp on the image plane. This condition is met if the image, object and lens plane intersect on a single line.

Scheimpflug Principle

If we are working with planes and multiple optical systems, it will be practically impossible to orient them parallel to the object's plane and viewing the same region of interest, especially if disparity is required (for instance to accommodate stereo vision). Therefore it is probable that these cameras or projectors will be oriented to view the plane from an angle. This means that the cameras DOF will not lie in-plane with the objects plane. Since we are working with large in-plane dimensions and small out-of-plane features, we will want to orient this plane of focus parallel with that of the object. Alternatively, we can set a very high aperture on the camera to increase the DOF, but diffraction will make the image appear soft and unsharp.

If the camera is oriented at an angle with the plane, we can orient the optical system with respect to the sensor such that the plane of sharp focus lies in parallel (and centered on) the object's plane. This method is called the Scheimpflug principle and is illustrated in Figure 4-6. This principle is well known and obvious in old view cameras, but less so in digital ones. This principle is incorporated in some lenses that are available for high end cameras and called *tilt*-lenses.



(a) Ordinary Camera

(b) Scheimpflug Principle

Figure 4-6: In (a) we see the basic optical geometry of a normal camera, in which each element is perpendicular to the lens axis. In (b) the Scheimpflug principle is illustrated, where the intersection (Scheimpflug line) defines the new orientation of the plane of sharp focus.

Time-of-Flight

A Time-of-Flight (TOF) camera is a scannerless Light Detection and Ranging (LIDAR) system that can capture the depth dimension of a projection on its planar 2-D sensor, similar to that of a normal camera. Its working is based on resolving the travel time of light rays that it emits and then captures upon their reflection based on their fixed speed [48]. These cameras do not account for inter-reflections on the surface, do not usually capture color information and capture little information at a time. Due to the high speed of light, depth information often has a low resolution and accuracy. Although these cameras are limited in these features, they can be successful in fusion with different approaches like with the regular stereoscopic approach, initializing the stereo-algorithms and giving it a disparity estimate, improving the eventual depth estimation and decreasing computation cost [49].

Light-Field

If a special micro-lens array is placed at the focal plane in front of a regular digital sensor, accounting for the lens array in the software makes it possible for it to yield '4-D light-field' information [50]. Such a camera is called a plenoptic or light-field camera. The eventual image that is projected on the sensor will resemble the image made with a normal camera. Instead of the light directly hitting one sensor, through the micro-lens array the light that would normally accumulate on one sensor element (one pixel), is segmented by angle of incidence. This does mean that for each pixel on a normal camera, you need at least dozens of pixels of this same point, to account for the different angles of incidence. This is shown in Figure 4-7.

The fact that you can account for the angle of incidence is the key feature in this approach. This means that you can account for very straight or sharp angles, which can lead to images with a near or distant focus point. Alternatively, you can account for rays coming from a certain side of the lens. If you only consider the rays coming from the sides, you effectively have two images that have a disparity as if taken side-by-side. This is very interesting for stereoscopic applications, because the image is made with the same camera and therefore a lot of parameters are fixed. The two viewpoints are then in perfect synchronization, static alignment and can share the exact same settings.

The plenoptic camera sacrifices a lot of spatial data in order to capture the light-field. A regular sensor with 10 megapixels that has a regular microlens array will only yield around

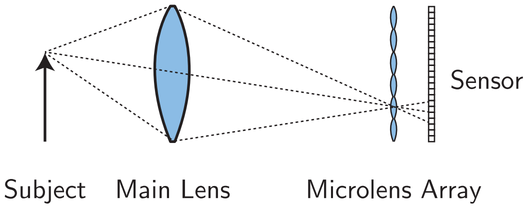


Figure 4-7: Schematic of a plenoptic camera. The subject is focused on the microlens array plane, which divides the converging rays onto the digital sensor.

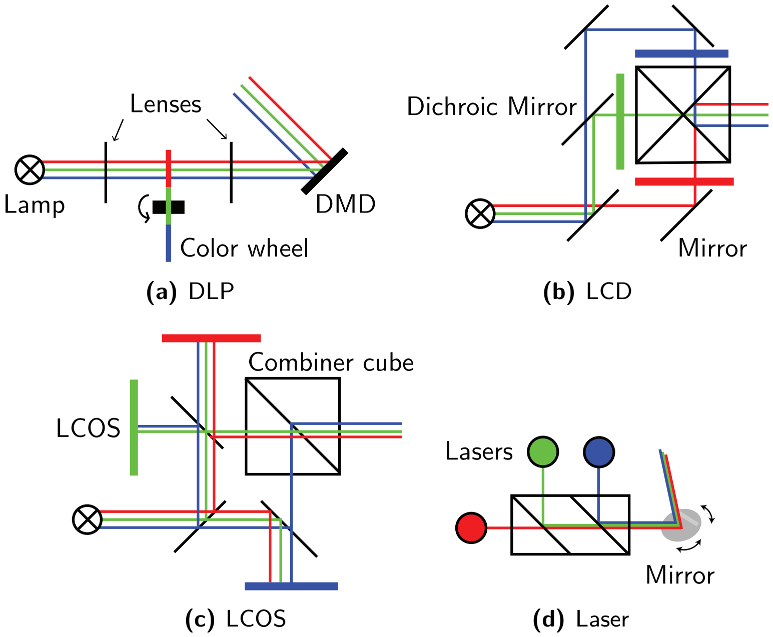
300x300 pixels, giving you barely 0.1 megapixels in return. With smart and expensive (and error prone) processing, sub-pixel (super resolution) accuracy and different optimizations are possible to extract a higher degree of spatial information [51]. Also, different focal distances can be combined to yield an extended depth of field.

4-2-2 Projectors

There are some active methods for depth perception that make use of structured light that is projected on the surface. This is historically done with gratings in front of a normal lamp, but more dynamic methods generally make use of digital projectors, which have recently become interestingly cheap mainstream products. These projectors have characteristics similar to that of an inverse camera, and should be taken into account.

The projection is done through a lens, which means that it is similar to the cameras optical system and the projection can exhibit certain artifacts like distortion and chromatic aberration, which can be accounted for with a straightforward inverse camera resectioning method [52]. Next to the similar properties like the optical system of a camera, it is limited in resolution, light output, and refresh rate. It can be important to take these parameters into account, and match these with the settings of the cameras capture system.

Important and often fixed parameters of a projector are the throw distance (distance from projector to plane) and the projected size. The technique used in projecting is also important. The most common projector principle is Liquid Crystal Display (LCD) as depicted in Figure 4-8(b). In this method, a lightbeam is split by dichroic mirrors, which are passed through an LCD screen before entering the combiner cube where the three color images are combined. In this case, the LCD screen modulates the light that passes through. In a Digital Light Processing (DLP) system, light is reflected on a Digital Micromirror Device (DMD) that contains thousands of acutated mirrors that modulate the light. In a one-DMD system (3-DMD systems exist but are costly and rare) the DMD can only modulate one color at a time. Color projection is achieved in this system, by very quickly projecting each color at a time. This is done by putting a rotating color wheel in front of the projector's lamp as depicted in Figure 4-8(a). This means that DLP devices will display a 'rainbow effect' if the observer is moving relative to the projection, or has a faster capture rate than the projector. The DLP system generally does have a superior dynamic range and efficiency over the LCD method. Another newer method, Liquid Crystal on Silicon (LCOS) could be seen as a hybrid between the two previous methods. It arranges the LCD screens as mirrors as depicted in Figure 4-8(c). Furthermore it should be noted that the LCD units themselves polarize the light, a feature inherent to the technique. This means that the usage of polarizing filters (like



(a) DLP **(b)** LCD **(c)** LCOS **(d)** Laser

Figure 4-8: Schematic of the working principle of different projection techniques.

the cross-polarization setup discussed in Section 2-2 to reduce surface specularity) in front of the projector or observer might not be possible. For our purposes, small projectors called 'pico-projectors' could be applicable. These projectors are very portable, efficient, and have a small throw distance that can be as small as 5 cm. This can be advantageous when we need very high resolution images where we have to move very close to our plane of interest.

Another projection method that makes use of lasers is depicted in Figure 4-8(d). It inherits the drawbacks that accompany lasers (see Section 4-2-3) but also gives it interesting features like a practically unlimited depth of field, although laser spot size does increase accordingly with distance. The particular laser projection that is depicted here combines the light from three spot-lasers with different wavelengths, which are reflected by a Micromechnical System (MEMS) mirror scanner. Due to the small nature it has an extremely low inertia allowing it to rotate very quickly around two axis, allowing for vertical and horizontal positioning of the laser dot. A bi-directional raster scan allows for the eye to see an area projection, while technically only one dot at a time is illuminated. This means that synchronization with potentially other time sensitive equipment is critical, although the display of the three colored channels is simultaneous. This projection method is not very popular because of its limited gamut, probably induced by the fact that (due to their nature) the spectral emission curves of lasers are very narrow in terms of wavelength bands.

4-2-3 Lasers

Like projectors, lasers can be used to emit structured light, albeit not as dynamic and versatile. In essence, the light emitted from a laser is very spatially coherent, yielding a beam that can be focused on a very tiny spot. This spot can be split through a diffraction grating to multiply the amount of emitted spots as successfully used by Microsoft's Kinect sensor. This spot can

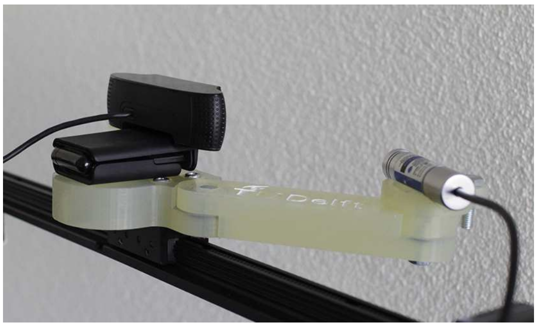


Figure 4-9: The laser-profilometric setup that was used as a pilot test. A laser projects a line onto a surface from an angle with the camera, which in turn can extract the profile cast by irregularities (depth) on the object of interest.

also be guided by lenses to accomodate the spot or change its shape, changing it into a line or cross laser for example. The use of lenses can induce distortion artifacts.

The usage of lasers is more static than the use of projectors, but using lasers can be beneficial through the ability of projecting very thin spots and lines with a very high light intensity, for use with triangulation and profilometry. The emitted light on a surface always has a certain thickness that limits the accuracy and resolution of the eventual system. Lasers also produce artifacts that are referred to and observed as speckle noise, an unpredictable interference pattern. In a pilot test during the making of this document, we used a laser line projection in a profilometric setup where the laser and camera translated along a surface. This setup is depicted in Figure 4-9. Instead of translating the entire setup, one could also incorporate rotating or translating mirrors to change the setups viewpoint. Laser profilometry is well established and often applied to scan depth. The main drawbacks that occur have to do with the defined width of the laser line or dot, intensity biases, optical calibration, the limited dynamic range and the fact that it does not let the camera capture the visual spectrum simultaneously.

Due to their coherence, lasers are also often employed for interferometry. In a typical Michelson interferometer, a beam of light is split in two by a half-silvered mirror. One beam travels down a path with a known distance that ends on a detector. The other path travels through the environment with an unknown distance, and reflects on it to end up at the detector. Due to their path distance difference and same speed, they end up at the detector that should be able to measure their phase shift. This phase shift, or interference, is the basic principle behind interferometry in general.

4-2-4 Confocal Microscopy

A particular setup that is able to reconstruct high resolution 3-D images is confocal microscopy, in which a very narrow aperture in front of the light source and observer eliminates all the light outside of the focal plane as illustrated in Figure 4-10. In this method, all the out-of-focus light is eliminated as it is not seen by the observer, and the depth resolution is therefore inversely related to the size of this focal plane. In order to sample over an area, one needs to either translate the pinholes, specimen, microscope or 'bend' the light rays using scanning mirrors. Using such a setup, one can reconstruct 3-D structures on a microscopic

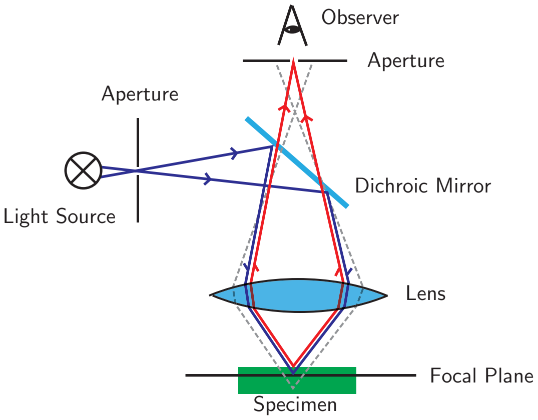


Figure 4-10: A rough schematic illustrating the working principle of a confocal microscope.

scale. The clear disadvantage of this method is that it is slow and is ineffective in capturing larger surfaces as that of an entire painting. A problem in using this method on a larger scale is that very small apertures lead to a large depth of field, diminishing the depth resolution. An example of the usage of the principle behind confocal microscopy on a larger scale is called 'synthetic aperture confocal imaging' [53]. In those cases the light source aperture is synthetic and induced by a projector, and a depth resolution of a few centimetres can be obtained using a normal camera.

4-3 Methods and Algorithms

The hardware that has previously been discussed can be arranged in a wide variety of methods that can accommodate 3-D measurements. Apart from these methods, one needs algorithms to take care of the processing of the data that have a significant influence on the system's characteristics and performance.

4-3-1 Camera Characterization

For any imaging system one can make use of techniques of camera enhancement where the performance of the device can corrected and/or enhanced. Corrections generally alleviate errors inherited by the nature of the hardware that was used while the performance or limitations imposed by this hardware can be extended by software enhancement.

We can estimate the geometric parameters of a camera system through a process called camera resectioning and calibration. Doing so, we can accurately relate the projected image on the cameras sensor to that of the world's coordinates. Because it is important to know your hardware's characteristics, such a process is not only useful for any camera system, but for some systems it is also a prerequisite, like in stereo vision. Geometric resectioning and color calibration, for stereo vision in particular, is described in detail in Section 5-2.

4-3-2 Stereo Vision

From a relative difference in viewpoints in two images induced by either the observer or the scene, one can match different points and construct a three dimensional map. The issue that addresses the matching of these points is called the correspondence problem.

When there is a relative difference in orientation and location of viewpoints in two points in time, one can perceive depth. This is called ‘depth from motion’. Because often the exact locations of those different viewpoints are not known, the accuracy is lower than when using two static and constrained viewpoints like in stereo vision. Dynamic scenes will pose further complications.

Fixing the baseline and the orientation of two cameras rigidly to enable stereopsis is therefore a good idea. The calibration of these cameras both separately and together must then be done with a high degree of accuracy, because a good image registration is essential for this method.

The image registration process of stereo vision is quite similar to that of stitching as discussed in Chapter 8. Instead of matching features from two different viewpoints to acquire the relative mapping function to distort the images, we now match features in order to obtain the disparity map that allows us to extract absolute depth information when the setup’s parameters are known.

The main drawback of stereo vision in the use of machine depth perception is that is can only triangulate positions where the data in the image is salient, where there are features that can be matched with enough certainty. Where there are good and identical features, stereo matching and triangulation can be highly accurate. Another important realization is that the data that is used to acquire the depth map in this method is actually RGB color data. Since this data is required as the texture of our depth map, we are in fact collecting all the data at once, eliminating the need for separate color capture and later matching and patching it onto the depth map.

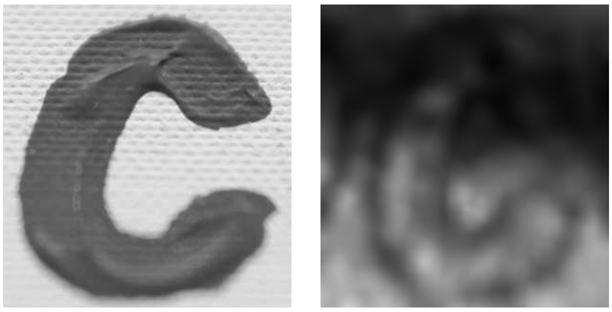
Stereo imaging is discussed in detail in Chapter 5.

4-3-3 Focus and Defocus

When a certain camera setup focuses, often certain features are in focus, and others are out of focus. This effect is stronger when the depth of field is small and the distance deviation in the scene is high. This means that the amount of (de)focus can reveal information about the depth of the scene. These cues are often not very strong or reliable, because there are more factors that can induce blur in an image like shadows or texture. The way focal blur behaves is also hard to model in mathematical terms. The way the lens distorts the objects that are out of focus is denoted as ‘bokeh’.

The relation between camera parameters and the amount of blur when moddelled with a Gaussian kernel is as follows [54].

$$D \approx \frac{F \cdot v_0}{v_0 - F - \sigma \cdot f} \tag{4-1}$$



(a) Regular image **(b)** Depth map

Figure 4-11: An illustration of depth from defocus. In (a), the bottom part of the image is out of focus. This fact can be estimated to produce a depth map as (b).

In this equation, D is the distance from lens to object, v_0 the distance between lens and focal point, F is the focal length, f the aperture number of the lens and σ the Gaussian standard deviation or blur size. This equation directly approximates the estimated blur from an absolute depth estimation.

There are several approaches to computing distance directly from images, as the blur size still has to be estimated. One could use sophisticated methods using edge detection, derivative maps and zero crossings to achieve this [55]. Another straightforward approach is convolving the image with two separate Gaussian kernels of different sizes, and with some simple arithmetics arrive at a blur map [56]. When implemented, this last method can yield a relative depth map as depicted in Figure 4-11.

4-3-4 Photometrics

Shading, highlights and gradients can directly reveal information about the surface topology. It is possible to extract cues from a single image alone, which is known as shape from shading [57]. These clues are subtle but can become useful when the illuminant is moved. This approach is very limited and has many obvious ill-posed problems. An approach that uses consecutive images of the same surface that is lit by a point source from different directions is called photometric stereo[58], as depicted in Figure 4-12. Through this method surface orientations are estimated, which yields a map containing the surface normals. In theory, the depth map can be computed from the normal map by integrating over the gradients. In practice, that integration will not perform well, as inevitable occlusions, discontinuities or the slightest miscalibration will produce errors that propagate through the entire depth map. In practice that integration is only proven to work well with systems that are highly standardized [59], which will pose limitations on the system that would otherwise be beneficial to this approach. For the sole purpose of deriving an estimate of the surface normals this method is remarkably straightforward and fast. The biggest limitations are that all illuminants are considered to be point sources and the surfaces are Lambertian, and even then occlusions are inevitable in practice as a finite amount of light directions is used. This limitation is overcome when the surface normals map is derived from the depth map, which yields results as good as the quality of the depth map.

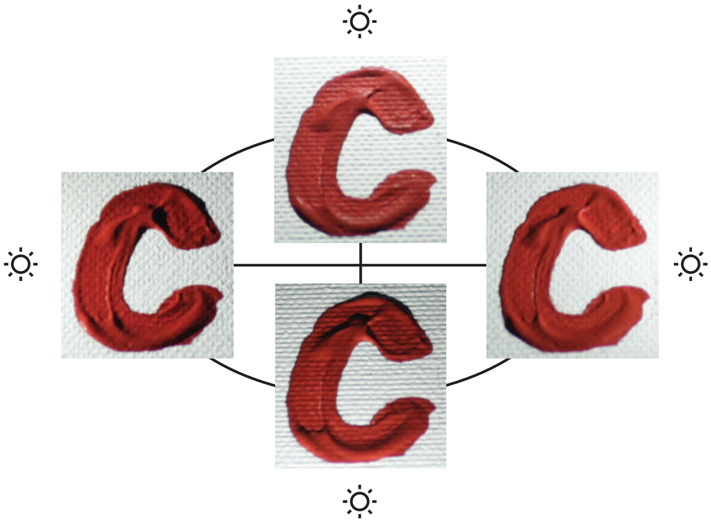


Figure 4-12: Surface with illumination from different distinct locations can be used together to yield information about the surface orientation. A polarization filter has been used to reduce specularity.

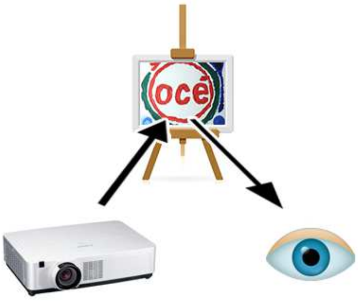


Figure 4-13: A typical structured light setup.

4-3-5 Structured Light

Triangulation or trigonometry is a basic and essential technique that is used for projecting structured data, whether it is emitted through a laser or a projector. In all methods, the camera and projector (sometimes multiple) have a certain disparity, and observe the scene from different angles. A typical structured light setup is depicted in Figure 4-13

The location of a laser dot can be triangulated using simple goniometry when the location of the source and observer is known. One can easily expand this idea by using a laser line, dramatically increasing the data acquisition speed. Especially with the projection of a laser line it is very intuitive to understand that a projected laser line reveals the profile of the surface. This basic idea can be extended even further to usage by projectors in the form of coded structured light, that has evolved into a field of its own through developments in digital projectors.

Using Coded Structured Light (CSL), one can capture large planes of depth data with the capture of just one or multiple images using basic mathematics, depending on the technique used. The amount of published techniques is quite staggering, some of which are classified

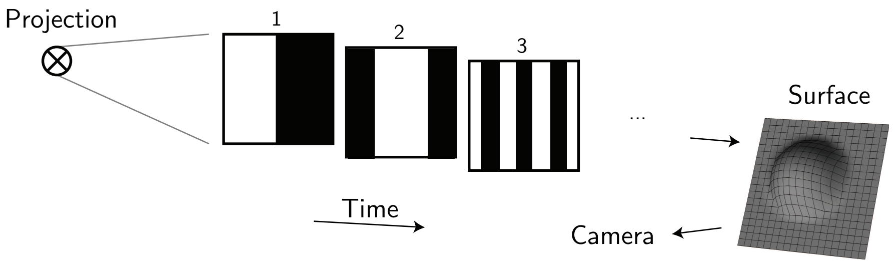


Figure 4-14: Consecutive projections on a surface in a typical structured light system. In this example, the first three figures are the first in sequence of a binary stripe pattern with a different spatial frequency. Adopted from [7].

Table 4-1: Classification of CSL methods. Adopted from [10].

<table>
<tr><td rowspan="2">Discrete</td><td>Spatial multiplexing</td><td>De Bruijn
Non formal
M-Array</td></tr>
<tr><td rowspan="2">Time multiplexing</td><td>Binary codes
N-ary codes
Shifting codes</td></tr>
<tr><td rowspan="3">Continuous</td><td>Single Phase Shifting (SPS)
Multiple Phase Shifting (MPS)</td></tr>
<tr><td>Frequency multiplexing</td><td>Single coding frequency</td></tr>
<tr><td>Spatial multiplexing</td><td>Grading</td></tr>
</table>

in Table 4-1. The most basic uses the projection of binary code with sequentially decreasing frequency as depicted in Figure 4-14. In this method, for every spatial coordinate in the image plane, a code is generated through multiple projections leading to its according depth value. This idea can be extended to include arrays of colors and stripes that change the properties of the system like making it faster or more robust [60, 10]. Hybrid methods are also used, most often together with fringe projection which is discussed in Section 4-3-6. Usage of such a hybrid method allows for sub-pixel accuracy. Otherwise, the spatial resolution is limited to that of the projector. Coded structured light can also aid in solving the correspondence problem [61] as discussed in Section 4-3-2.

One of the problems with structured light projection arises with surfaces that are specular. In those cases projected structures might be reflected and projected indirectly on unintended locations. When projecting a repetitive image, a linear diffuser can be placed between the projector and object, which significantly reduces specularities [62]. Alternatively, specularities can be addressed by using more randomized texture projection encoding [63] although this inevitably leads to a decrease in per-frame spatial resolution.

4-3-6 Fringe Projection

The technique of fringe projection for extracting 3-D surface information has become one of the most active research areas in optical metrology [8]. Affordable and versatile projectors and cameras can be used in a variety of different setups to yield good results with sub-pixel

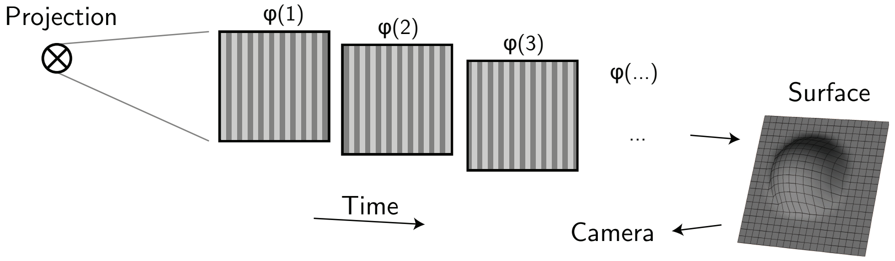


Figure 4-15: Consecutive projections on a surface in a typical fringe projection system. In this example, the figures of the sequential projections are fringes that are slightly shifted in phase with each projection until a full cycle has been obtained.

accuracy for a wide range of applications. Technically, fringe projection can be classified as an active form of structured light projection. It is alternatively called a phase shifting technique and its computations are denoted as fringe (pattern) analysis. Its methodology is not so obvious and intuitive as the previously discussed methods of coded structured light, and we will therefore elaborate on it. It can also be used together with standard coded structured light, where the fringe projection serves to acquire the finest degree of accuracy.

The typical image acquisition schematic for this system is depicted in Figure 4-15.

The phase shifted images can be described with formula (4-2), where I' is the average intensity and I'' the intensity modulation. The goniometric entry determines the amount of phase shift and ϕ the amount of induced phase shift of each successive fringe pattern[9].

$$I_r(x,y) = I'(x,y) + I''(x,y) \cdot \cos\{\varphi_r(x,y) + \phi\} \tag{4-2}$$

With multiple projections, one can solve for the phase shift to yield the following general equation for an arbitrary amount of phase images.

$$\varphi_r(x,y) = tan^{-1}(\frac{\sum_{n=1}^{N} I_n(x,y) \cdot \sin(\frac{2\pi n}{N})}{\sum_{n=1}^{N} I_n(x,y) \cdot \cos(\frac{2\pi n}{N})}) \tag{4-3}$$

The minimal amount of phases (and thus captured images) needed for this method is three. In that case, (4-3) can be rewritten and simplified as follows.

$$\varphi_r(N=3) = tan^{-1}(\sqrt{3} \cdot \frac{I_1 - I_3}{2 \cdot I_2 - I_1 - I_3}) \tag{4-4}$$

We could then also combine these three images to give us the average intensity $I'(x,y)$, intensity modulation $I''(x,y)$ and data modulation $\kappa(x,y)$. In the case of three phase shifts, we would obtain equations (4-5) and (4-6) to yield $\kappa(x,y)$ through equation (4-7).

$$I' = \frac{I_1 + I_3 + I_2}{3} \tag{4-5}$$

$$I'' = \frac{\{3 \cdot (I_1 - I_3)^2 + (2 \cdot I_2 - I_1 - I_3)^2\}^{0.5}}{3} \tag{4-6}$$

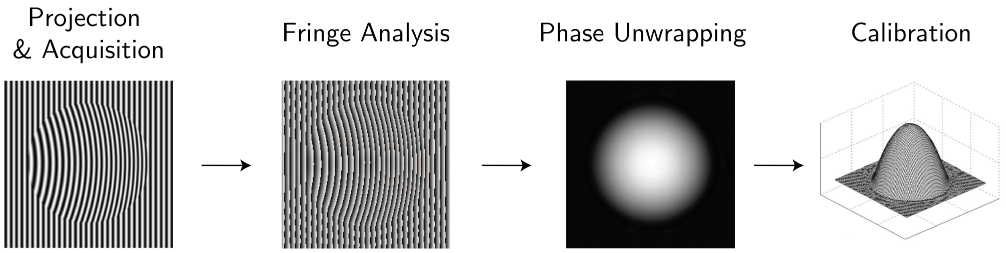


Figure 4-16: Flow chart for the fringe projection technique. The phase unwrapping image is here illustrated as a depth map. Adopted from [8], images from [9].

$$\kappa = \frac{I''}{I'} \tag{4-7}$$

The data modulation κ acts as a quality metric for the wrapped phase image φ_r. At each pixel, a value of 1.0 means a high signal-to-noise ratio (SNR), while lower values approaching 0 are candidates for omission of the data at that point.

After acquisition, the schematic of the data processing is depicted in Figure 4-16. The previous fringe processing functions will yield us a phase map. This phase map is in a wrapped state, which means that the surface depths are enclosed within a $\pm 2\pi$ range. These now have to be unwrapped to yield the actual (unwrapped) depth data. Phase unwrapping is a nontrivial task and several disparate solutions exist. Wrong wraps will produce errors that can propagate and effect the entire phase map. Adequately accounting for low SNR's and using enough phase steps can help prevent these artifacts.

Another nontrivial task in fringe projection that is often overlooked, is the nessicity of calibrating the camera *and* projector for both geometry *and* color. Especially in the case of using computer generated fringes, one needs to account for the nonlinearity of the gamma of the camera and projector [64], which would otherwise lead to significant phase measurement errors resulting in harmonics on the surface topology.

Because fringe projection allows us to compute the average intensity as in equation (4-5), we also capture the color of the object simultaneously and unambiguously. A special case where the three minimum fringes are shifted in the three color channels [65] does not account for this. On the contrary, that approach does allow one-shot fringe projection capture, but assumes a homogeneous surface color.

Fringe projection is discussed in detail in Chapter 6.

4-3-7 Projection Moiré

Before digital projectors were mainstream equipment, multiple gratings were often used to project repetitive structured light. The method that was often used to extract the topography is called projection moiré or moiré topography. In optics, moiré refers to a beat pattern produced between two gratings of equal spacing. With an appropriate setup, this phenomenon can be used to derive topographical data and is technically the oldest form of fringe projection. In a basic setup, a grating is projected onto a surface, and observed through a reference

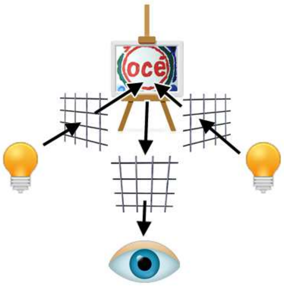


Figure 4-17: Practical setup of the projection Moiré technique.

grating. These gratings interfere and form moiré fringe contour patterns that are dependent on the surface topology. Setups often include moving (Ronchi) gratings and multiple projectors [66]. A typical setup is depicted in Figure 4-17.

Projection moiré is more computationally expensive, ambiguous, often has moving parts, is less accurate and sensitive than the fringe projection methods [67].

4-3-8 Sonography

Another non-invasive imaging technique that is dependent on sound waves rather than light is sonography, and ultrasound-based imaging technique derived from sonar (sound navigation and ranging). The high (ultra) frequency is used because it provides a higher resolution due to the smaller wavelength, but the depth penetration through the medium decreases. The emitted ultrasound reflects on transitions between media with different characteristics, notably the acoustic impedance which is largely dependent on the density. For best performance, the emission crystal should be coupled to match the medium using a film. The relative impedance from air to any material is in the order of 1:1000. Besides problems with coupling, because of the small resolution (in industrial application often single point sample; echolocation) the acquisition speed is very slow. That fact that we cannot use any other medium than air for our application therefore limits the possibility of using sonography.

4-3-9 Radar

Instead of LIDAR, that (by definition) uses the electromagnetic wavelength of (nearly) visible light to estimate range, one could also use the very long wavelengths in the frequency band of radio waves. The well known term for that method is radar (radio detection and ranging). This method has no advantages over LIDAR for our purposes due to its long wavelength and single-point scan nature that limits accuracy and speed.

4-3-10 Hybrid Methods

As noted in Section 3-2-12, the human fuses depth information observed through different cues and sensors, even when they conflict. Applying this same sensor fusion can also be done in machine vision, combining the methods that have been previously discussed. Stereo matching can be improved in both computing cost and accuracy through the usage of an extra Time-of-Flight camera [49]. Stereo vision can also be used together with projection methods like structured light or lasers. Using multiple cameras for these methods can yield just another viewpoint, reducing occlusions and increasing redundancies, which would rather be a dual system than a hybrid system. Many hybrid systems use multiple cameras, logically leading to an implementation that involves stereo matching.

A true hybrid system uses multiple methods that are different in nature, like the usage of projecting an uniquely encoded pattern on a surface, which is seen by both cameras, and then matches the unique codes in both cameras to solve the stereo matching correspondence problem [68]. This method is different from normal structured light, where one would relate the observed encoding back to the projected image of the projector. It is also different from normal stereo matching, where generally parts of their two normal 'pictures' of a scene are related to each other. The latter method will yield a very high stereo matching accuracy, and is actually often used to benchmark 'normal' stereo matching algorithms. This method will be elaborated upon in Chapter 7. Another approach [69] uses a projector in stereo matching to project 'unstructured light', a random pattern projected on a surface used to overcome the fact that many man-made objects lack dense surface texture, and when they do, it is often repetitive and can lead to erroneous correspondence. In this way, it induces its own random texture on a surface. Yet another method improves stereo matching through fusing it with relative motion of the object of interest and the camera, called 'spacetime stereo' [70]. It improves correspondence accuracy through not only the normal matching in space while staying a match over time through dynamic motion.

4-4 Related Work

Much work has been done already in the digitization of 3-D structures. In order to investigate the feasibility and practical aspects of the different acquisition methods discussed in this chapter, we can look at similar projects that have been undertaken and published.

One of the largest and possibly most prestigious practical 3-D work is the *Digital Michelangelo Project* [1] by Stanford University. It digitized the *David* statue, with a surface area of 19 m^2. They used the line-laser projection method to yield excellent results although the acquisition alone took a total of 1080 man-hours. The scanner could rotate around two directions and translate around three, and the contraption would be moved around the statue manually for different angles. The digitization by IBM of the *Florentine Pieta* by Michelangelo was of similar complexity [27]. It is unique in the way of its redundancy. It employed a hybrid of photometry, coded structured light as well as irregular laser dot projection, and six cameras to capture the texture. First the lasers and structured light were projected for alignment and mesh acquisition. Then photometry was used to refine this map, and the texture was applied.

Donatello's *Maddelena* statue was digitized[28] using structured light, a basic single projector and single camera setup called OPTO3D. The setup was unautomated and placed on a static tripod. The statue was digitized with a 'medium' resolution of 125 µm, and specific details were imposed onto that with the highest (75 µm) resolution, although the real depth accuracy itself is not stated. The depth data looks very good, although it shows significant occlusions around areas of depth transitions.

Structured light was used to digitize the *Vittoria Alata* statue [29]. The setup was a basic projector and a camera with a fixed disparity mounted on a regular tripod. They used around 500 different views which they stitched together to yield the digitized version of this 2 m high marble statue. Alignment time of all these views took 23 days using third party software. The surface accuracy was reported to be 90 µm to 400 µm for 90% for the surface. Another method that uses structured light for heritage objects makes use of color coding [71].

Illustrations on different innovative calibration panels are used for the calibration of the projector and the camera. Using their method, the Minerva statue was digitized. It reports problems with depth of field limitations that arise with coded structured light. They tested their method by measuring depth deviations on a flat surface, resulting in a value of 300 µm. More case studies are published in a single paper that make use of structured light [72], in which they digitize a statue, bust and painting from the Louvre. They used three commercial scanners with a feature accuracy of 50 µm. Matching of different point clouds was one with a least squares method. Remarkably they employed texture mapping separately with normal photographic equipment, while their scanners use simple gray coded structured light together with fringe projection. This method should allow for the simultaneous capture of the texture color. The in-plane spatial resolution of the painting was stated to be 60 µm with a 3 µm depth resolution. Nothing was said about the actual effective accuracy of this resolution. They tested several software suits to stitch different views together in a coherent map, but no single piece of software was satisfactory; instead, they used multiple.

A conoscopic micro-profilometric scanner is also used for scanning paintings [73]. Essentially this is a simple interferometric static point laser scanner. They report a 6 µm accuracy with a depth of field of 4 mm and a distance of 40 mm from the panel. Since the probe is static, a linear translational stage was made around it. Acquisition speed is unsurprisingly slow, around 300 points per second. Published results were in very high detail.

Apart from separate case studies, overviews that summarize different methods for 3-D scanning of cultural heritage objects have been made [74, 75, 76, 77].

4-5 Performance Evalutation

In a paper investigating laser scanner accuracy [78] in which a dozen high end commercially available laser scanners are tested, the following remark is made.

> If a scanner shows 'better' results than another one, this does not necessarily mean that it is the better instrument for a certain task different from the ones performed in our tests!

It would be impractical to compare different performance metrics stated by different sources in literature, because of our case specificity. Therefore we will give an overview and summary of the usage of different methods on a per-approach basis. In order to be able to compare the approaches that suit our intentions, we have to first state our requirements and then compare the methods to it that we have previously investigated.

4-6 Performance Metrics

(i) **Size**
This constitutes the width (x), height (y) and depth (z) dimensions of the planar object. The system should be able to efficiently digitize planes with x and y dimensions between 10 cm to 10 cm and 1 m to 1 m. The out-of-plane (z) depth deviation should be no more than 1 cm.

(ii) **Resolution**
The in-plane (x,y) resolution should be related to what we want and can print. Assuming that our printers resolution is not limited, we limit ourselves to the resolving limit of the human eye. For the average healthy human eye, the visual acuity is 77 cycles/° at its most sensitive [79] (the fovea). Relating this to digital pixels, we need two pixels to define a cycle, so the pixel spacing should be at least $1/77 * 2 = 6.5 \cdot 10^{-3}$pixel/°. If we then assume the eyes of the observer to be 0.75 meter away from the plane, the resolution should be $\text{tand}(6.5 \cdot 10^{-3}) \cdot 0.75 = 85\mu\text{m}$, which equals to $25.4/85 \cdot 10^{-3} = 298$ dpi.
The human visual system does exhibit hyperacuity [80] (super-resolution) beyond the theoretical resolving power based on the receptors on the fovea, we will aim for a spatial resolution of around 300 dpi, a value well established in the printing industry.

(iii) **Spatial Accuracy**
The accuracy should be such that we can obtain the effective resolution as stated in (ii). This means that at this resolution we preferably want a sampling efficiency[1] of 100%. In practice, such a high efficiency is not realizable and we will aim for a sampling efficiency of at least 85%, as recommended in the Technical Guidelines for Digitizing Culutral Heritage Materials [82] by the Federal Agencies Digitization Initiative Still Image Working Group (FADGI).

(iv) **Color Accuracy**
Because the final reproductions are in full color, the color encoding should be accurate. Similarly to the spatial requirements, our color encoding and accuracy should follow the FADGI requirements [82].

(v) **Speed**
The system should preferably be as fast and efficient as possible (in terms of points per second). Paintings that are of interest are often of high value and are displayed to public during daytime. Including overhead, the typical total capture time frame should not exceed one night.

(vi) **Manual Post-Processing**
The post-processing time should be able to be highly automated, and take no less than one hour of manual work. This includes all potential manual tasks like stitching, removal of noise, errors, occlusions, artifacts, etcetera.

(vii) **Invasivity**
The scanner should be non-invasive. This means that the system should in no way have

[1] Sampling efficiency is a metric that describes how effective or 'sharp' the image has been sampled [81].

any lasting influence on the object, other than would be the case in normal conditions. Therefore, a non-contact method is necessary. Active optical methods that use radiation in any wavelength (λ) are also bound within limits. Dutch guidelines for radiation on paintings state [83] that the temperature difference between the surface a light radiates on and its environment should not exceed 3°C, and IR and especially UV radiation should be avoided. The brightness of museum lighting for sensitive objects is generally accepted to be around 50 lux.

(viii) **Portability**
The proposed system should be portable, as the scanning for our test objects will be mostly used on location.

(ix) **Cost**
Due to budget limitations, hardware expenditure is to be kept within few thousand euros.

4-7 Comparison of Methods

An overview of the major fields of 3-D imaging methods that have been discussed in this chapter is benchmarked in relation to our requirements and depicted in Table 4-2.

Table 4-2: Performance of 3-D imaging classes.

Class	Approach	Section	Size & Resolution	Spatial Accuracy	Color Accuracy	Cost	Speed	Manual Post-Processing	Non-Invasivity	Portability
Time-of-Flight		4-2-1	-	-	-	-	+	+	o	+
Light-Field		4-2-1	-	+	+	o	-	+	+	+
Laser	Interferometry	4-2-3	o	+	o	-	+	+	o	+
Laser	Trigonometry	4-2-3	+	+	-	+	+	+	o	+
Stereo	Dense	4-3-2	+	-	+	+	+	-	+	+
Stereo	Sparse	4-3-2	-	+	+	+	+	-	+	+
Depth from Motion		4-3-2	+	-	+	+	+	o	+	+
(De)focus		4-3-3	o	-	+	+	+	-	+	+
Photometric		4-3-4	+	-	+	+	+	-	o	+
Coded Structured Light	Binarized	4-3-5	+	+	o	+	o	+	o	+
Coded Structured Light	Textured	4-3-5	+	+	-	+	+	+	o	+
Fringe Projection		4-3-6	+	+	o	+	+	+	o	+
Moiré		4-3-7	+	o	-	+	+	+	o	+

The time-of-flight method is extremely fast and portable with a very dense in-plane data yield without occlusions. This method does not capture color and our own brief depth estimation tests exhibited errors and biases on changing albedo and reflectivity that can not easily be accounted for.

The light-field method can be used as a stereo vision method with a high amount of simultaneous different baselines. The employed micro-lens array will effectively reduce the spatial resolution severely.

An approach with using laser interferometry can be very fast but is highly expensive. It also shows strange artifacts and a bias on inhomogeneous surface intensity as our tests have displayed using a Faro 120 laser interferometric scanner. A trigonometric approach is feasible with the use of a line laser, but depth-of-field and dynamic range are a concern. Moreover, no RGB map is captured.

Stereo vision is a feasible solution. Cameras have to be accurately calibrated (in color and geometry) and depth-of-field should be accounted for in the setup. Capture rates can be high, but images without much saliency are a concern, and erroneous correspondences would have to be filtered out. Depth from motion can be implemented similar to stereo vision, and can yield more baseline images to correspond with. The estimation of the movement is essential, and therefore stereo vision with a rigid baseline would have a preference. With neutral lighting, the color reproduction could be optimal and simultaneous.

Using the cameras depth of field with depth from defocus is not very feasible as blur estimation itself is very coarse, and it would need very salient features, leading to a sparse map.

When using photometry, the system will have to be very standardized and rigid. Because of complications that arise from integrating a surface normal map to a depth map, this method is likely to yield errors that may cascade through the imagery. Moreover, one assumes a Lambertian surface, which is hardly the case when a varnish is applied.

Coded structured light and fringe projection are good candidates as they yield a dense and accurate map with a high resolution. Fringe projection is not limited to the resolution of the projector and therefore has a higher accuracy. Artifacts are often observed with the usage of these methods, arising from the sensitive calibration of the projector and camera, of both their color and their geometry. Their simultaneous color reproduction is fair, but highly dependent on that emitted by the projector, which is in practice never spectrally neutral. Environmental light is preferably dimmed during capture. Moiré projection has no obvious advantages over structured light.

4-8 Conclusion

The best candidates for the topographic scanning of surfaces such as paintings are trigoniometric laser scanning, stereo vision and structured light projection. The feasibility of these methods is confirmed through the fact that these are the main methods used successfully in related work on paintings as described in Section 4-4.

From the overview in Table 4-2, we can see that no single method can ideally meet our requirements. Although these methods are indeed feasible in practice, a system without concessions that exactly suits our requirements can be constructed if multiple methods are imposed simultaneously. The likely candidate is a combination of stereo vision and structured light projection due to their nature, while they are different in the way they solve the correspondence problem, in a passive and active method respectively. This means that they can solve each others problems because their shortcomings are naturally different. For example, stereo vision alleviates '2π' discontinuity problems with the phase unwrapping while structured light can solve photo-consistency problems in stereo vision. The observation from multiple viewpoints would also reduce occlusions.

We will now investigate separately stereo vision and structured light (specifically fringe projection) in respectively Chapters 5 and 6. Then we will proceed by fusing these two methods in Chapter 7.

Chapter 5

Stereo Vision

Most stereo algorithms make assumptions about known and good camera calibration and epipolar geometry, because this is the best understood part of stereo vision[84]. The input for most algorithms is a pair of rectified images. When both images are rectified, they are both projected onto a common image plane, which simplifies the problem of finding matching points between two images, because each feature now shares one same dimension value (epipolar assumption). In a two dimensional image, one searches for a corresponding point or area in the other dimension that matches. The value difference of the found matching locations within these two images is called their disparity map. A simplified view of this process is illustrated in Figure 5-1. Repeating this process for all features in the image will construct the two dimensional disparity map, from which the depth map can be constructed.

Stereo matching algorithms can be divided into sparse and dense methods. Sparse methods find salient features in both images and match these. Their speed and accuracy is high, but their resolution is dependent on the amount of good features. A particularly popular algorithm for describing features is the scale-invariant feature transform (SIFT) algorithm that allows keypoints like edges, corners or gradients to be matched to each other [85].

Dense matching tries to match every part of one image to the other, often in a window based approach, locally or globally while they minimize a certain cost function. One particularly popular approach is the usage of an adaptive window[86]. Common pixel based matching costs include sum of squared differences (SSD) and sum of absolute differences (SAD) approaches. In practice, most algorithms follow four steps, matching cost computation and cost aggregation, followed by disparity computation/optimization and its refinement. Performances of each of these particular methods is often highly dependent on the scene and its parameters[84].

Apart from its straightforward setup, stereo vision (in both machine vision and in humans) does have several limitations, one of which is the smoothness assumption that nearby pixels in the image plane are nearby in reality. This is practically never the case in an arbitrary environment, although in our particular case of flat and often smooth surfaces, this assumption is more valid.

Moreover, one can often construct a cost mapping in which the quality (or certainty) of the disparity map is displayed. Susceptible regions can then be accounted for. Another

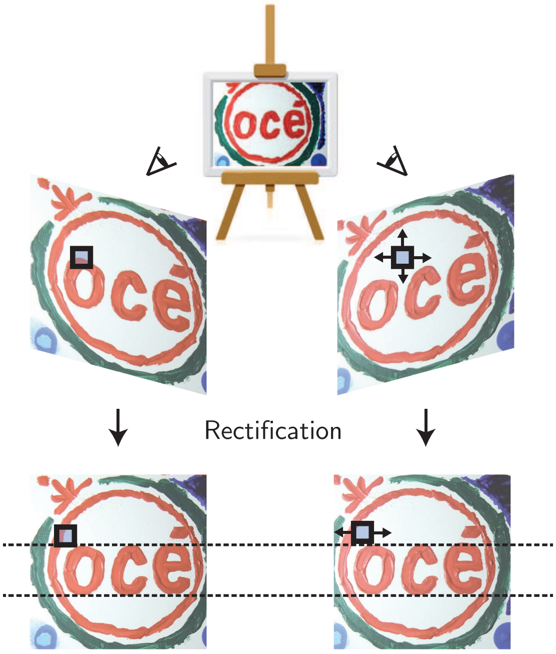


Figure 5-1: With stereo vision, rectifying the two images captured by both cameras reduces the dimension in which the disparity occurs.

problem is that the cameras are not exactly identical and the scenes might reflect light with a different brightness. This is known as the problem of photo-consistency. This means that the color in both cameras should be calibrated as well. Furthermore, when no features are present one cannot accurately match two regions, especially given the uncertainty imposed by photo-consistency due to the different cameras. Such problems could in theory be relieved by employing a hybrid with an active light projection method.

5-1 Color Calibration

A good color reproduction is generally desirable for any camera that has the ability to capture color, especially for objective purposes like reproductions and research. This is particularly important for stereo cameras, where their degree of color accuracy is not only important in relation to the actual color they observe, but they should also be accurate in relation to each other. In most stereo matching algorithms, one or more pixels in one camera will be related to the other, in which their respective pixel values are taken into account. This means that color accuracy and its calibration can be of importance in order to achieve a valid result.

So when an accurate color reproduction is desirable, one needs a camera that can capture a good representation of the spectral reflectance of each of the colors in the scene. This representation is often done in the tristimulus fashion as described in Section 4-2-1. When the colors are then encoded in the RGB space, a color profile is often appended that describes how the actual RGB values relate to displayed colors. The RGB colorspace by itself is thus perceptually nonuniform. Due to the quantization, this often standardized (ICC) color profile puts limitations on the size of the 'spectrum' of colors it can capture. The range of colors that

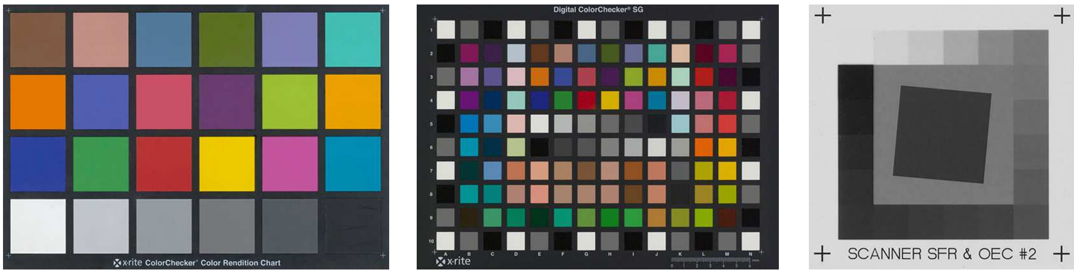


Figure 5-2: Three different targets used for camera calibration and performance benchmarking. Color-patch based targets (a) and (b) are industry standards used for color calibration and benchmarking, and target (c) is a slanted edge target that can be used for characterization of sharpness.

it can successfully capture is called the color gamut, as previously depicted in Figure 3-2. It should be noted that the color filters often put constraints on the quality of its capture that cannot be simply overcome through digital correction.

An RGB camera can be calibrated through the usage of color calibration targets as depicted in Figure 5-2. These targets have a wide collection of patches with different spectral reflectances. The captured RGB values for each patch are then matched to their reference values, and from this data, an input ICC profile can be constructed. Such a unique profile is called a custom profile, and can be converted to a standardized and common ones like sRGB, Adobe (1998) or the more modern ECI-RGB as advised by the European Color Initiative (ECI). Then, this profile can be applied to the image so that the colors will be displayed correctly. For printing, the output ICC profile of the printer should also be constructed (through printing and measuring of colored patches).

Construction of ICC profiles is a non-trivial task, and considered more of an art than a science, as there are several approaches towards constructing such a profile. Construction of a custom ICC profile can be done through open source software like Argyll CMS. The practical application of color calibration is depicted in Figure 5-3. The assignment of a custom ICC profile to an image, and conversion to a standardized one is illustrated in Figure 5-4.

Color reproduction performance can be measured by taking an average ΔE. This is done through relating the captured LAB values (after ICC profiling) to the reference LAB values. The (CIE-)LAB color space is considered to be perceptually uniform. The ΔE metric is essentially the perceptual difference between two colors as observed by the human eye. The average ΔE is a single number that averages the ΔE values of each of the patches in the calibration target.

As was the case with custom ICC profiling, a correct conversion from RGB to the LAB color space is more complex than it seems. In this conversion, the specific ICC profile has to be taken in account, as well as the illuminant (D50,D65, . . .) gamma (1.8, 2.2, . . .) and observer angle (2°, 10°, . . .).

To make color calibration and its performance measurements even more complicated, the ΔE metric knows different methods of computation that result in significantly different values. The way of computation has changed over the decades due to the fact that the LAB color space turned out to be not as perceptually uniform as intended, which modern ΔE computations account for. Most common and simple is the '1976' method, which is just the Euclidean

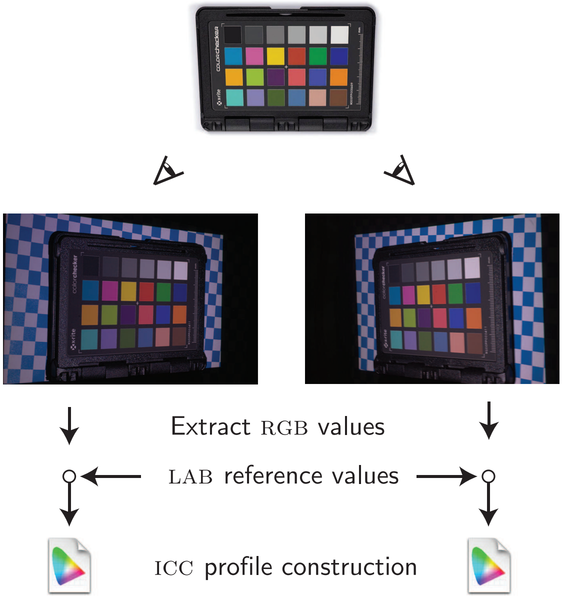


Figure 5-3: Workflow for color calibration in a stereo setup. A separate (custom) ICC profile is generated for each camera, using the extracted RGB values of the patches that are related to their reference values in the LAB color space.

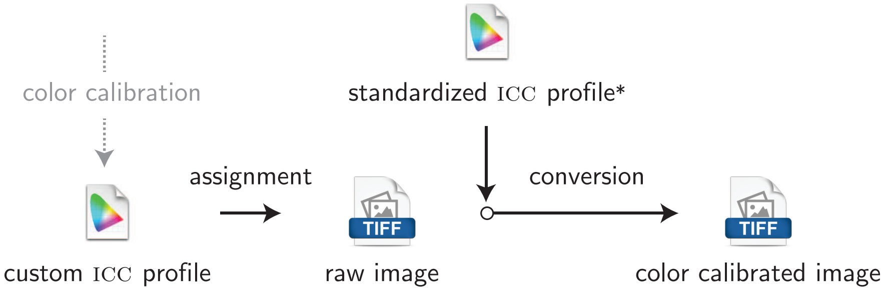


Figure 5-4: After the application of a colorprofile to an image, it is wise to convert it to an accepted and standardized profile. The assignment of a profile is nothing more than appending metadata, while a conversion requires recomputation of the actual RGB values. (*=SRGB, ECIRGB, etc.)

distance between two LAB values. More accurate, complex and advised by the Commission Internationale de l'Éclairage (CIE) is the '2000' version of the metric[87]. To make a clear distinction, values obtained with the latter computation will be referred to as ΔE-2000.

5-2 Spatial Calibration

We can estimate the geometric parameters of a camera system through a process called camera resectioning and calibration. Doing so, we can accurately relate the projected image on the cameras sensor to that of the world's coordinates. We can classify these techniques into two categories, photogrammetric and self-calibration. The former uses an object whos geometry in 3-D space is known, while the latter uses movement of the camera itself through a scene.

Photogrammetric calibration is well established and there are several open-source toolboxes that have this functionality implemented like the 'Camera Calibration Toolbox' for MATLAB or OPENCV for C++. These methods extract intrinsic and extrinsic parameters. Intrinsic parameters are those imposed by the camera model itself, encompassing focal length, principal point, skew coefficient and distortions. Extrinsic parameters are those that describe the relation between camera and environment in terms of rotations and translations. Several techniques that estimate these parameters are available, like a classic two-stage approach[88] or a well established flexible one that is very straightforward to implement[89]. From several captures in different positions and orientations of a checkered plane of known dimensions, all the parameters can be extracted. In practice, it is important to realize that these parameters might change during operation due to automated hardware changes like auto-focus or zooming. Fixing all of these parameters and turning automated functionality to manual is therefore advisable.

Stereo calibration is essentially the same as a single camera calibration. The difference is that in order to relate the cameras with respect to one another, the same fiducials have to be selected in the same real-world positions. That way, the relation between the views of the planes defines their relative position. Performing stereo calibration, the uncertainties of the intrinsic parameters decrease relatively to a single calibration due to the fact that the global stereo optimization is performed over a minimal set of unknown parameters. The workflow for (stereo) calibration of the well established method that uses the flat checkerboard plane as reference is illustrated in Figure 5-5.

5-2-1 Parameter Estimation

In order to estimate the parameters, we can use the pinhole camera model. Equation 5-1 relates the perspective projection from real-world coordinates X_i, Y_i and Z_i (with respect to the camera reference frame) to the image plane (pixel) coordinates ui and v_i. In this equation, f is the focal length, k_u and k_v the scale factors, u_0 and v_0 the projection of the optical center and ρ the skewness. The matrix that contains the intrinsic parameters is denoted as matrix A.

$$\begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} \begin{pmatrix} f \cdot k_u & \rho & u_0 \\ 0 & f \cdot k_v & v_0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{bmatrix} X_i \\ Y_i \\ Z_i \end{bmatrix} \tag{5-1}$$

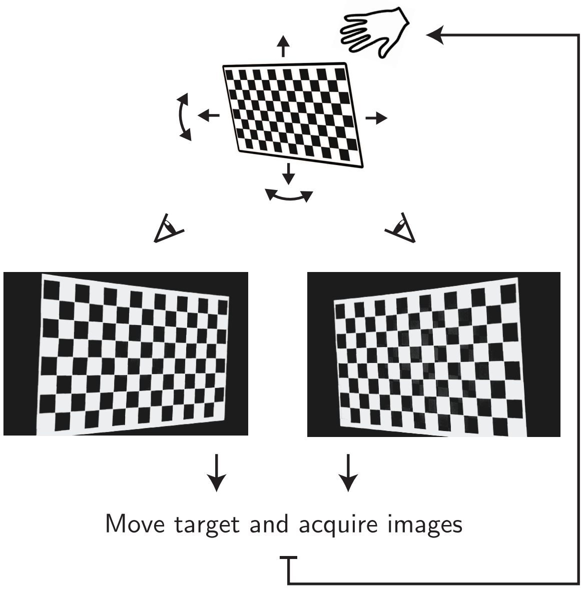


Figure 5-5: Workflow of spatial calibration of a typical stereo camera setup. A target (checkerboard) is captured at multiple locations in space to allow for unambiguous spatial calibration through ray triangulation.

When these intrinsic parameters have been estimated, we can compute the projection ray vector. The previously mentioned internal camera parameters that have to be estimated can be divided into 9 degrees of freedom. Accurate estimation of this multitude of parameters can only be done through the capture of multiple distinct views that solve the equations.

The extrinsic parameters relate the real world 3-D coordinates to the 3-D camera coordinates, and can be denoted as rotation matrix R and translation vector T. Knowing all these parameters allows us to relate any pixel on the sensor to a 'light beam' vector coming into or out of the cameras optical center. Doing so for multiple cameras in a stereo setup allows us to triangulate a 3-D position, once two matching pixel locations are found.

5-3 Algorithms

With the use of camera calibration we can accurately triangulate the position of an object projected on both cameras sensors. The main obstacle in stereo matching is the correspondence problem, relating one point that is projected on one cameras sensor, to the other. Epipolar geometry or homography estimation can help us make this problem easier by rejecting (or not looking for) the matching of points that cannot possibly be related to each other. Image rectification as illustrated in Figure 5-1 makes the correspondence problem easier and is often applied in dense stereo matching. Stereo matching can be roughly divided into two categories: dense and sparse. Dense matching means that for each pixel in one image, we attempt to match it to the other, while sparse matching generally just looks for salient keypoints and relate them to salient keypoints in the other.

5-3-1 Dense Matching

If we wish to capture a depth map in which each pixel in one image is matched to another, dense matching is used. Generally dense matching is done along an epipolar line, which simplifies the density of the matching to one dimension. Dense matching therefore generally is time consuming. A classic example of dense matching is the usage of area matching, in which an small area in one image is related to that in another along the epipolar line. This relation is given a score computed with for example the cross-correlation or sum of squared difference. The best match is then stored, and its relative disparity in pixels is saved. A depth map can be computed from the disparity map because it is inversely proportional.

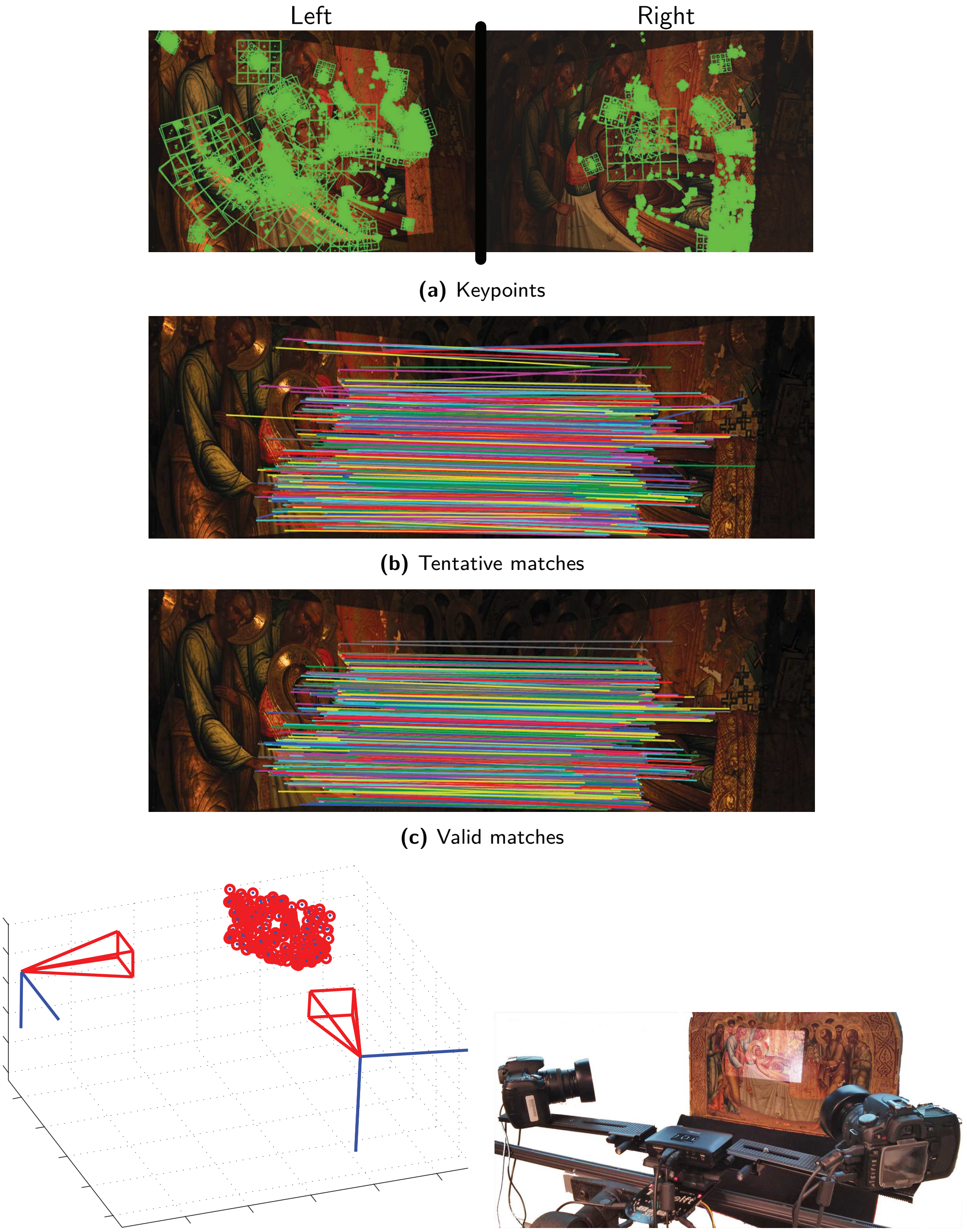
Dense matching has a few obvious drawbacks. Although it will attempt to match each region in one image to another, it might not always find the correct value, often because either the data is not salient enough, or there are ambiguous areas (repetitive textures for example). It also generally poses the smoothness assumption, and might not handle occlusions well. In the case of paintings or planar surfaces dense matching will certainly not perform well on surfaces with homogeneous color, especially when the color intensity that is captured varies slightly between the cameras. In such cases, we will require a more active vision approach, like projecting our own induced patterns with a projector.

5-3-2 Sparse Matching

As stated previously, one popular stereo matching algorithm that uses sparse matching is the SIFT algorithm that allows extracted keypoints like edges, corners or gradients to be matched to each other [85]. The advantages are that only keypoints are extracted that are distinct enough to match. We can see an example of extracted keypoints in a side-by-side image in Figure 5-6(a). This happens in both of the images, and the keypoints that have a close resemblance are then paired, and these tentative matches are illustrated in Figure 5-6(b). Like dense matching, we can exclude pairs that do not fit our geometry, which leaves us with the pairs depicted in Figure 5-6(c). In this example, we have found a few hundred pairs that match very well. These pairs can then be triangulated, which is displayed in Figure 5-6(d). Conveniently, there is an open-source SIFT toolbox available called VL-SIFT which greatly simplifies the implementation of the keypoint extraction and matching.

5-4 Matching Using Scene Encoding

The previously discussed stereo vision implementations have centered around solving the correspondence problem by the matching of intensity values between two gray scaled or colored images. Solving the correspondence problem in this way will neither provide reliable nor high accuracy solutions in regions where distinct features are scarce in the light intensity maps. Such problems can easily occur in paintings, where it is possible that large surfaces are sometimes painted without much distinct texture or color differences. But stereo matching does not limit itself to matching such intensity values. One can impose its own labeling of an image through, for example, a projector. The form of structured light projected that is ideally suited for this is binary or gray-code projection, discussed in Section 4-3-5. That method will uniquely label each pixel from the image projected by the projector, and illuminate this on



(a) Keypoints

(b) Tentative matches

(c) Valid matches

(d) Triangulated matches

(e) Setup

Figure 5-6: A sequence depicting the steps in sparse stereo matching. In (a) salient keypoints are detected that are good and unique enough for matching. Then in (b) the keypoints are matched with those extracted from the other camera with a close resemblance. In (c) the keypoints are rejected that do not match our homography. After that, we perform ray tracing and intersection with both cameras, that results in a point cloud containing high quality triangulated matches in (d).

a point in the scene. This point is then also uniquely labeled, irrespective of the observer's viewpoint. Therefore, we can solve the correspondence problem by matching these labels in multiple images with a high accuracy [68]. The problem with these particular coded structured light methods is that their performance is highly constrained by the projector's performance and features (especially its resolution). The usage of the fringe projection as a structured light technique suffers from less dependence on the projector's features and performance. Therefore, fringe projection is interesting to use for this implementation, although it is not naturally suited for this method due to its repetitive and relative nature (whereas the binary or gray-code technique labels in an absolute fashion). A method to use fringe projection as a labeling or image encoding technique will be discussed in Chapter 5.

5-5 Conclusion

Stereo vision techniques and its matching corresponding points can yield a very high accuracy and resolution in determining 3-D locations of observed points. However, these methods have difficulties coping with regions in both images that are not salient. This will result in the inability of triangulating these points, or doing so with a high uncertainty. This is especially a problem when reconstructing a painting with regions that are painted with a single color without many distinct features. For parts like these it would be advantageous to use a passive method that imposes its own distinct texture on parts of the scene that are ambiguous to our stereo algorithms. We have seen in Section 4-7 that coded structured light, in particular fringe projection fits those requirements and is well suited for our application. Therefore we will explore its potential in Chapter 6, and discuss its fusion with stereo vision in Chapter 7.

Chapter 6

Fringe Projection

Fringe projection has been previously introduced in Section 4-3-6, and it is a subclass of the (coded) structured light projection method. It makes use of the projection of a few repetitive sinusoidal signals with each a shifted phase. This is different from using binary or gray coding to project a unique pattern on a scene, so that the encoded 'light beams' as observed by the camera can be traced back into a specific pixel emitted out of the projector. Then each observed encoded part in the environment can be easily triangulated.

The latter could be seen as the projection of an absolute encoding of the scene, whereas fringe projection encodes the pattern it projects in a fashion relative to the previous fringe. Fringe projection provides many advantages over more common structured light methods, like speed, accuracy and its independence of the resolution of the projector. This comes at the cost of increased difficulty in implementation, limitations, and handling of errors native to the method.

While in theory the working principle is simple (see Figure 6-1), practical problems can be overcome by making good choices in the implementation of the design.

6-1 Design Considerations

An important contributor to the performance of fringe projection is the choice of the parameters of the fringes themselves, and the method and nature of their respective projection.

6-1-1 Fringe Parameters

The projected fringes have the same characteristics as that of the common sinusoid, like wavelength (frequency), amplitude (modulation) and offset.

A small wavelength means that the fringes will be spaced very close to each other. In any case, the cameras will have to clearly distinguish these fringes. Since the projector and camera are inherently spaced apart and aimed at a common plane with a different angle, some parts

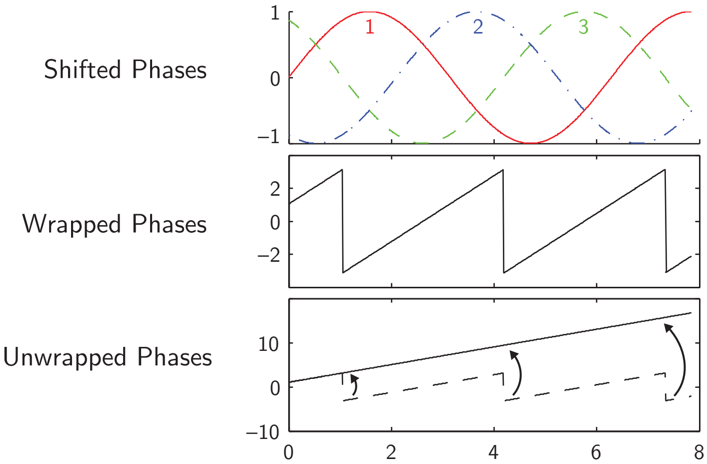


Figure 6-1: Simple illustration of the technique behind fringe projection. Fringes (sinusoids) with specific phase shifts are captured and then wrapped using an appropriate goniometric operation. The wrapped phase signal can then be unwrapped, and related to a depth deviation.

of the projection might not be visible to the camera. Such occluded area's are not necessarily a problem as smart fringe unwrapping algorithms will be able to 'wrap around' them. But this becomes an increasing problem when one fringe or more is occluded by the camera, in which case our algorithms have no possibility to detect that it has missed an occluded fringe. This is because fringes are repetitive and individually indistinguishable in nature (as opposed to other more absolute encoded structured light methods like gray or binary coding). On the contrary, a large wavelength will decrease the signal-to-noise ratio. A small signal to noise ratio will eventually saturate subtle depth differences with the larger error deviation.

The amplitude and offset are important in the sense that it might be possible that the digital projector is unable to display intensity values in their extrema, like the brightest white (8-bit value '255') or darkest black (8-bit value '0'). This is possibly due to the display settings and conversions of the Graphical Processing Unit (GPU) of the host computer to the independent internal display management of the projector, or some other limiting value. Failing to account for this can result in the sinusoidal to be clipped in its extrema. A high modulation of the fringe is advantageous to the eventual signal-to-noise ratio when wrapping the images. On the other hand, a very high modulation will result in a highly varying intensity over the image, which might contain both black and white colors. Black colors that are illuminated by the minimum of the fringe and white colors that are illuminated by the maximum of a fringe, will have a very high intensity difference, requiring a very high dynamic range.

6-1-2 Amount of Fringes

The general equation that governs the principle behind this method, can be solved using a different amount of fringes n_f, with a minimum of three (the equation to phase wrap these images is stated in Equation (6-1)). For example in the case of three fringes, the fringes on the images $I_1 \dots I_3$ are shifted with increments of 120° to yield a full phase 'revolution'. For some applications, this number is convenient as this is exactly the amount of color channels in a digital camera. This means that this method can be used by capturing only a single image, on which a red, green and blue fringe are projected that are spaced 120° apart. However, that method assumes a homogeneous surface color intensity, an assumption that conflicts with our requirements.

The minimum amount of required fringes is three, and this effectively solves the equation and will yield a depth map. In practice however, more fringes are projected, commonly four (Equation (6-2) or six (Equation (6-4)). The reason is that in practice, projected fringes are not ideal sinusoidal fringes, often because the camera and projector's 'input/output gamma' is not modeled (see Section (6-3-1)). The effects of using more fringes than theoretically necessary is twofold; it increases the amount of captures required (reduces speed), but it reduces the dependence on the projection and capture of a perfect sinusoidal signal. An increasing number of fringes will have corresponding harmonics, sometimes of multiple frequencies, but their amplitude will be smaller [90] and will reduce the final deviation. One remarkably elegant solution that significantly reduces these harmonics is a double-step solution [91]. The double-three (or six) step algorithm is stated in Equation (6-4). In this case, it projects two times a three-step pattern, with one of those patterns shifted by half a phase-step. For example, in the latter case, the fringe phases will be shifted as $(0°, 120°, 270°)$ and then by $(0° + 0°, 120° + 60°, 270° + 60°)$. The two patterns will be individually wrapped and unwrapped, after which they are averaged. Elegantly, the harmonics from the first pattern are canceled out by the harmonics of the second pattern. In practice this works very well, although double the amount of captures are needed. One study [92] found that the single-three step method with different accurately modeled gamma corrections still performed worse than the double-three step method without correcting for the gamma. Still, in any case it is worth characterizing the 'gamma' as best as possible as accounting for the gamma is a mathematical procedure that will not negatively influence performance.

Below the Phase Shifting Profilometry (PSP) equations are given for different amounts of fringes n_f (or images) per period. Their result yields the wrapped phase φ.

The common three-step PSP method [93].

$$\varphi(n_f = 3) = \tan^{-1}(\sqrt{3} \cdot \frac{I_1 - I_3}{2 \cdot I_2 - I_1 - I_3}) \tag{6-1}$$

The common four-step PSP method [93].

$$\varphi(n_f = 4) = \tan^{-1}(\frac{I_4 - I_2}{I_1 - I_3}) \tag{6-2}$$

The five-step PSP was developed by Hariharan et al. [94].

$$\varphi(n_f = 5) = \tan^{-1}(\frac{2 \cdot (I_2 - I_4)}{2 \cdot I_3 - I_5 - I_1}) \tag{6-3}$$

The double-three step (six-step) PSP algorithm developed by Huang et al. [91].

$$\varphi(n_f = 6) = \frac{\tan^{-1}(\sqrt{3} \cdot \frac{I_1 - I_5}{2 \cdot I_3 - I_1 - I_5}) + \tan^{-1}(\sqrt{3} \cdot \frac{I_2 - I_6}{2 \cdot I_4 - I_2 - I_6})}{2} \tag{6-4}$$

The seven-step PSP developed by De Groot [95].

$$\varphi(n_f = 7) = \tan^{-1}(\frac{I_1 - 7 \cdot (I_3 - I_5) - I_7}{4 \cdot I_2 - 8 \cdot I_4 + 4 \cdot I_6}) \tag{6-5}$$

These functions are all different due to the amount of fringes, are the functions that derive their average intensity I', intensity modulation I'' and data modulation κ. The generic form of these three formulas is stated as follows. Note that in Equation (6-7) the variables *num* and *denum* stand for respectively the numerator and denumerator of their corresponding arctangent operation for the computation of φ.

$$I'(n_f) = \frac{1}{n_f} \cdot \sum_{i=1}^{n_f} I_i \tag{6-6}$$

$$I''(n_f) = \frac{1}{n_f} \cdot \sqrt{num^2 + denum^2} \tag{6-7}$$

$$\kappa = \frac{I''}{I'} \tag{6-8}$$

The average intensity map I' averages each phase shifted fringe, leading to an image that yields the intensity of the scene as would have been seen by a normal RGB camera without fringes and a normal illuminant. Intensity and data modulation I'' and κ can be used as thresholds, where parts of the phase map with values too low can be rejected.

6-1-3 Focused and Defocused Patterns

When attempting to project a fringe rather than a binary pattern, we assume we project a continuous and smooth sine. When using digital systems, the luminance of the pixels in the projection that make up the sines will be discretized to a certain value. The result is that in a sharply focused projection, this discretization can be made out by the observer, making the fringe quantized instead of smooth. Therefore, smoothing the sine by defocusing can be advantageous, although defocusing does inherently decrease the fringe's amplitude. Ideally, the fringe's amplitude is just large enough to fall within the dynamic range of the camera, which in turn will discretize the data. Defocusing works because a sine (fringe) that is convoluted with any kernel, will produce a sine wave of the same frequency (convolution theorem). This also reduces the need for a projector with a large resolution (as is the case with binary- or gray-code projection). In order to retain the strongest fringe amplitude while using defocusing, one can use a strongly defocused binary pattern [96], resulting in a fringe pattern.

6-1-4 Digital and Analog Projectors

The modern definition of a 'projector' is one of the digital variety. Illuminating a scene with structured light can also be done through analog devices, as was often done using Moiré projection or early structured light implementations. The main difference in practice is that analog projectors (like a slide or overhead projector) are more static and will require mechanical parts in order to change the structure of the emitted light, whereas the digital variety typically has a refresh rate of 60 Hz. The main advantage of the analog projectors is that one could easily use a custom illuminant, for instance a spectrally neutral light, instead of using the RGB primaries with highly irregular spectral curve that is found in digital projectors. One disadvantage is that the spatial accuracy of repeatability of the digital projector is very high or even perfect, while such accuracy is hard to achieve in mechanical systems.

A digital projector often has an offset in projecting the RGB primaries, projecting each primary next to the other (no exact 'color overlap'). In the case of the analog projector, there will typically be no offset in these RGB primaries when good 'slides' are used, other than chromatic aberrations induced by the optics.

6-1-5 Multiple Observers

The common setup of this method involves one camera and one projector. One could also use multiple cameras that observe the scene from different viewpoints. Then the algorithms that were applied to the images from the first camera can also be applied on the others. The final results (depth maps) form each image can be fused together to yield a more complete map without occlusions due to the different viewpoints. The problem is that each camera will exhibit the same errors that are induced by an inaccurate calibration or gamma estimation (see Section 6-3-1). A more hybrid approach where different cameras work together with fringe projection is discussed in Section 6-6.

6-2 Spatial Calibration

The calibration of cameras has been discussed in Section 5-2, and is well established. Calibration methods for the projector system are less well established and includes a number of specific methods. More generally, we can also use the inverse camera calibration model [97], adopting its versatility and flexibility. This is because a camera receives rays as if projected from a surface, while a projector projects these rays onto a surface. The camera should be calibrated first, after which the projector can be calibrated whilst, for instance, projecting the checkerboard pattern.

For the camera calibration, a physical reference chart (commonly a checkerboard) can be used with fixed dimensions. With multiple views of this same checkerboard of known geometry, the camera model can be solved. In the case of fringe projection, we have a camera and a projector that are fixed in space. Since multiple views are necessary, we will have to move a plane in space in front of the camera and projector, on which a known image is projected. We trace a ray coming out of the projector and into the camera, and from that compute the real world coordinate of the surface on which is projected. In theory, such a triangulation sounds trivial, but we have no extact a priori knowledge of the intrinsic and extrinsic parameters of the setup. It is important to keep in mind that the rays are bent by the optical system of both the projector and the camera, posing further challenges. Moreover, multiple views (board positions) are necessary to fix the system and provide a unique solution.

6-2-1 Calibration Methods

In the previously proposed setup where projector and camera are pointed at a plane that moves between different views, one can still not derive any calibration parameter. In order to do that, one also needs to know the exact position of the plane on which is projected. When the intersection of the projection-ray to plane is known in real world coordinates, the rays can simply be traced back as straight vectors into the projector (once emitted, light in a homogeneous medium does not bend over humanly distances). The bending does indeed occur in-plane over the image due to the projector's optics.

As with projector calibration in general, several solutions exist. This can be done by placing fiducials with known positions on a plane [98][1].

One method projects coded markers [99], so that fiducials can be distinguished, a feature needed when the entire projection is not in view. This method too assumes a pre-calibrated camera rig.

Alternatively, a mechanical solution can be adopted, where the reference plane is mechanically displaced very accurately [7, 100]. Another method called steerable projector calibration [101] uses steered a pan-tilt mirror to change projection poses. The projected target is in fact in black and white, which means the physical fiducials have to be separately illuminated by ambient light due to the black projection.

The previous approaches are remarkable, as in most cases the camera still has to be either calibrated separately or requires an extra operation before the projector calibration.

All calibration procedures generally use either a single channel or gray-scale image, as hues are not important in spatial calibration, except when cheap or wide angle lenses are used that feature a lot of chromatic aberration. Therefore, one can easily devise a method in which the (physical) camera calibration occurs in one channel, while in another channel one can extract the projected image. One can then use the fiducials from the physical (camera) calibration in order to derive the orientation and location of the plane on which the projection takes place. Another important advantage here is that through camera calibration, these same physical fiducials are actually checked in quality through error analysis, native to camera calibration (how well they eventually fit the cameras parameters).

6-2-2 Combined Camera-Projector Calibration

The main challenge in the previously proposed combined camera-projector calibration is that the channels have to be separated effectively. The camera has to distinguish the color emitted as the checkerboard target by the projector in a separate channel, while this channel is insensitive to the color as emitted by the physical calibration target. Moreover, in another channel, the camera has to distinguish the physical checkerboard, while being insensitive to that of the projected checkerboard. This poses a challenge to the RGB color filters of both the projector, the camera, as well as the (in case of a printed physical target) CMYK colors of the physical checkerboard. Ideally, one should characterize the transmission spectrum of all

[1] As proposed by the creators of the Projector Calibration Toolbox for MATLAB, which is unsuited for practical use (incomplete, flawed).

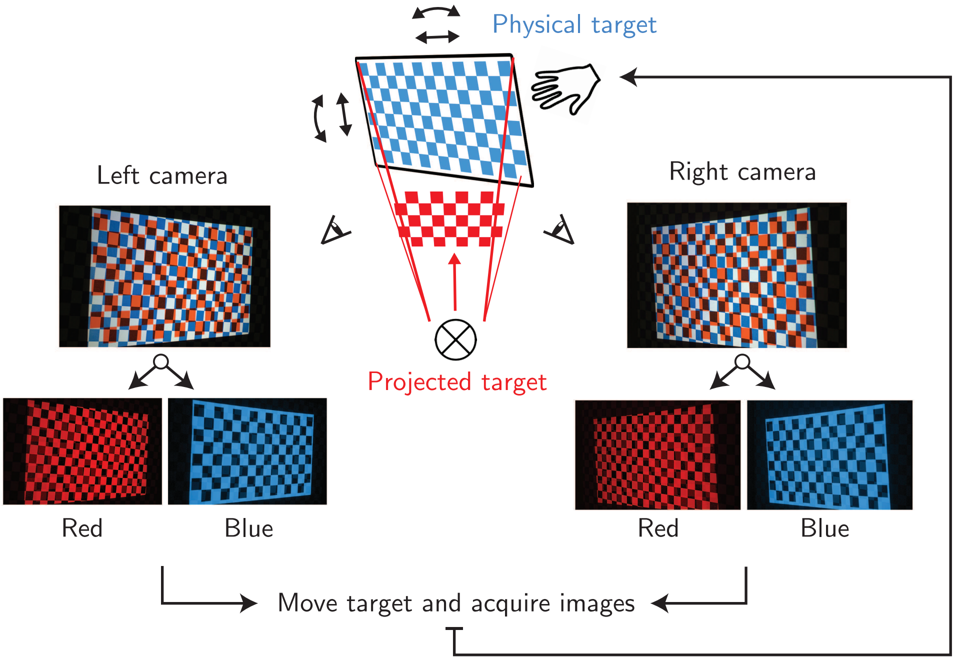


Figure 6-2: Proposed workflow for spatially calibrating a projector and stereo camera setup. A red checkerboard image is projected onto a physical cyan colored checkerboard. Color channel separation results in acquisition of the physical and the projected checkerboard. Multiple images of the target in different positions allows for unambiguous spatial calibration.

of these color filters, as well as the absorption of the printer's color spectrum. From there, one can select different colors that distinguish themselves the most in each channel.

In lack of multispectral sensors and emitters, a pragmatic solution was devised in a trial by error approach. First, a range of colors were printed on a printer by taking into account the ICC output profile of the printer as provided by the manufacturer. These colors were then captured by the camera in the light of the (brightest) illumination of the projector. These images were captured by the camera, and the colors that were indistinguishable from white in certain colors were selected. From there on, different colors were projected from the projector onto a white surface. Images taken by the camera were then assessed on how well these colors in turn were indistinguishable in which channel. Four our particular setup, the best performance was found when the projected pattern was of a red color (on a white background) onto a cyan colored checkerboard (with a white background). The projection can then be found in the blue channel, and the physical checkers can be immediately extracted from the red channel.

Finally, the workflow of the proposed combined (stereo) camera-projector calibration is illustrated in Figure 6-2.

After all the calibration targets have been captured, an optimization procedure estimates the calibration parameters. It then analyzes with sub-pixel accuracy how well the checkerboard positions and projection rays fit the estimated data, after which bad images (blurred for example) can be rejected. From this data, the extrinsic parameters can be visualized as depicted in Figure 6-3. The internal intrinsic parameters of each optical system can also be reviewed graphically as shown in Figure 6-4.

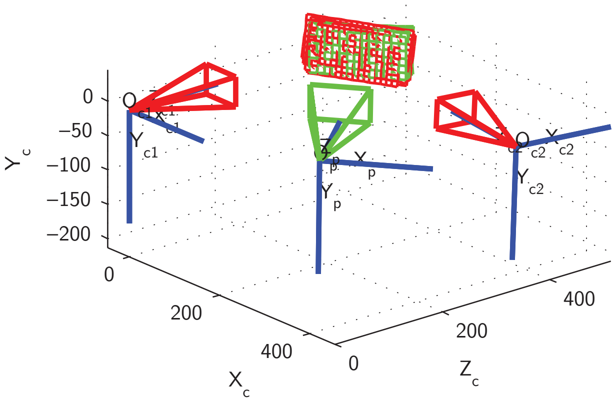
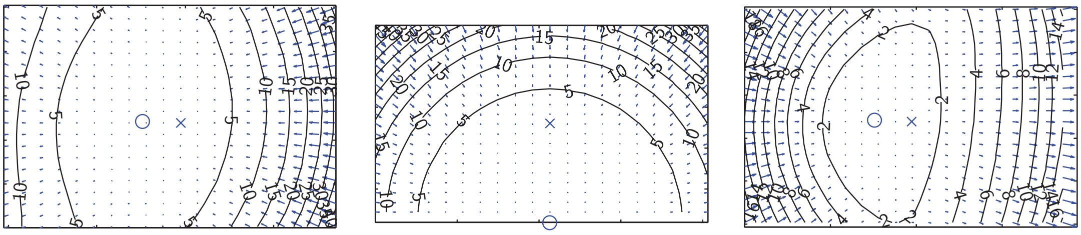


Figure 6-3: Illustration of the stereo-cameras (in red) and projector (in green) setup as the extrinsics, showing their respective locations in space, view pyramids and the checkerboards they were aimed at or emitted during calibration.



(a) Left Camera **(b)** Projector **(c)** Right Camera

Figure 6-4: Illustration of the distortion model (intrinsics) of the left (a) and right (c) camera and projector (b). These images display the displacement of the pixels within the sensor's surface through the ISO lines and arrows indicating the distortion shape. The principle point is displayed by a circle. Note the principle point of the projector is in the very bottom, this is because our projector projects slightly upward.

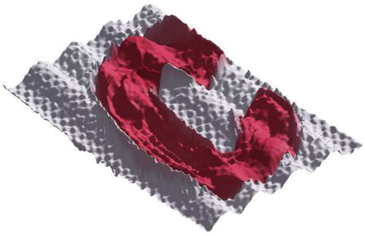


Figure 6-5: A three dimensional render of a canvas with a painted character 'C' that exhibits a strong harmonic depth bias due to inaccurate gamma correction.

6-3 Color Calibration

As we will not use the projector to project specific colors, its calibration is of much lesser importance than that of the cameras color capture since we will not project specific hues and saturations. The projector will project monochrome fringes (although built up by the RGB primaries). From this light, we will eventually compute the color of the surface as captured by the cameras as well. Therefore we will not calibrate the projector in the sense of constructing an ICC profile for it, but we will measure its characteristics as an illuminant, which will be incorporated in the construction of the ICC profile of the cameras. Therefore we have to measure the color temperature of the projector when it projects a neutral color (monochromatic, as done when projecting fringes). One important realization here is that when digital projectors are used, the illuminant is not spectrally neutral. This means that the illuminant cannot be considered a blackbody that emits radiation correlated with a certain (color) temperature, in accordance with Planck's law. In practice however, the approximated color temperature is still usable, although in our case metamerism becomes an increasing problem (Section 2-4-1). Ideally, the color temperature should be the same as when the colored object was conceived by the artist. A pitfall is that the color intensity and temperature of a projector can change in relation to their operating time, often due to warming up of the illuminant. The light response will become stable when the light is warmed up and stabilized. This is of less importance to projectors that use lasers as illuminantion source. This phenomenon also occurs in cameras, as photosensitive sensors have a dependence on temperature, but this is accounted for in the cameras hardware.

6-3-1 Gamma Correction

The projection of inaccurate sinusoidal waveforms is known to cause significant errors to the phase measurement [8]. These errors in the pattern are repetitive and present in each wave, therefore appear in the depth map as harmonic depth biases as illustrated in Figure 6-5.

These errors are due to the non-linearity of the gamma in both the camera and the projector. In the case of the projector, this means that the binary input results in a disproportional output in intensity. Therefore, when the binary input is a sinusoidal sequence, the intensity of the projected waveform will indeed be a waveform, albeit warped by the non-linearity. The largest contributor of this non-linearity is the purposefully implemented gamma correction in these devices. Gamma correction is commonly conceived to have its origin in outdated

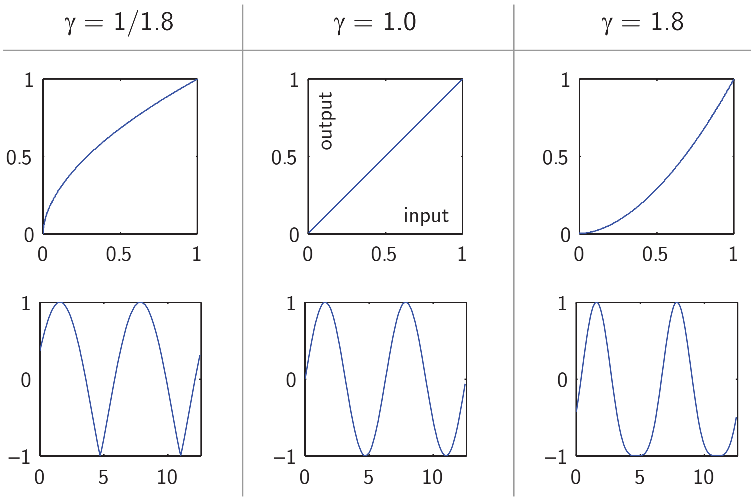


Figure 6-6: Example of three input/output gamma curves that are related with three different values of γ. Underneath each of the gamma transfer functions an illustration is given of what happens to a normal sinusoidal signal if a gamma function is applied to it.

Cathode Ray Tube (CRT) monitors, yet gamma correction still exists in any display or camera device. The human eye generally perceives images with fairly neutral gamma (around $gamma = 1.1$) as having the best image quality and naturalness [102]. This agrees with the conception of the gamma parameter in color profiles having its origin in CRT monitors.

The common gamma (γ) function for input signal V_{in} and output signal V_{out} (scaled within the range of $[0, 1]$) is stated in Equation 6-9.

$$V_{out} = {V_{in}}^{\gamma} \tag{6-9}$$

6-3-2 Camera and Projector Gamma

The unwrapping algorithms in gamma projection naturally assume their input to be originated from pure sinusoids. When a projector is given a fringe image to project, the internal gamma function will distort this signal accordingly in its projected brightness on the surface. The camera will then capture this fringe, and will distort it in turn by its own gamma correction. The signs of the gamma value of the projector and the camera are inversed, as one converts digital data to brightness, and the other captures brightness and converts this to data. It would be a misconception to assume that these two inverted gamma values alleviate each other, even when they are identical. This is because the brightness range of the projected light will never lay exactly on the minimum and maximum range of that of the camera, as there is often an offset and a scaling factor. Three different gamma transfer functions and three corresponding distorted sinusoidal signals are depicted in Figure 6-6.

6-3-3 Modeling Gamma Correction

There are several strategies that allow accounting for the gamma values. One particularly simple, effective and popular method is using a look-up table (LUT), and is one of the best

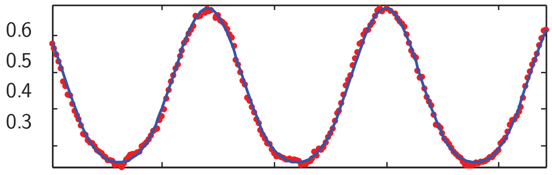


Figure 6-7: Observed projected sinusoidal intensity values (dotted) and estimated function (line) through modeling and parameter estimation accounting for device dependent gamma.

performing methods [103]. In that method, one unwraps several fringes projected on a plane, and then looks at the straightness of the line of the wrapped phase. For every phase value in the $[-\pi, \pi)$ range, the deviation from a perfectly straight line will be saved in a LUT. This LUT can then later be applied to every wrapped image. One important problem with this method is that the multiple fringe images can be averaged to yield a normal RGB image, but the LUT can only be applied to the wrapped phase image. The result will be that banding can be seen in the RGB image, a sign of incorrect accounting for gamma. Moreover, when the projector is defocused, the gamma value will automatically decrease (the principle behind the defocussed pattern technique, Section 6-1-3). The LUT will then prove to be invalid.

In any case, it is wisest to correct for gamma as much as possible before using any specific strategy. Most cameras have the ability to output raw un-demosaiced raw data, where no gamma is inherited from a specific profile. Alternatively, one can convert any non-raw image to an RGB profile that has a linear gamma (also called an L* profile), like eciRGBv2. For projectors and GPU's that cannot directly control the gamma, it still has to be estimated.

After obtaining a projected image with the camera, gamma estimation can be done quickly and accurately through straightforward parameter estimation with a model as in Equation 6-10. When an observed sinusoid y is projected on pixel array x, we want to estimate the parameters $c_n n = 1..8$ to yield y_{mdl} such that $|y_{mdl}^2 - y^2|$ is minimized. The parameters can be reliably estimated using a trust-region reflective algorithm [104], where one can also set bound constraints to each parameter. This is natively implemented in MATLAB. In Equation 6-10, c_6 and c_8 are the projector and camera gamma values respectively. Values c_1 and c_7 are the cameras scaling and offset values, while c_2 and c_5 are those internally of the projector. Parameters c_3 and c_4 are the frequency and offset of the specific sine to be reconstructed and are of lesser importance.

$$y_{mdl} = (c_1 \cdot (c_2 \cdot \sin(c_3 \cdot x + c_4) + c_5)^{c_6} + c_7)^{c_8} \tag{6-10}$$

One result of this parameter implementation is depicted in Figure 6-7. We can see that the observed values are very accurately modelled, especially because we have the a priori knowledge that the cameras gamma is around 1 and can be constrained around it. However, even with a good gamma estimate, harmonic errors will still be visible in practice (albeit with a low modulation). We can account for this by using 'self-calibrating' methods as the double step algorithms described in Section 6-1-2.

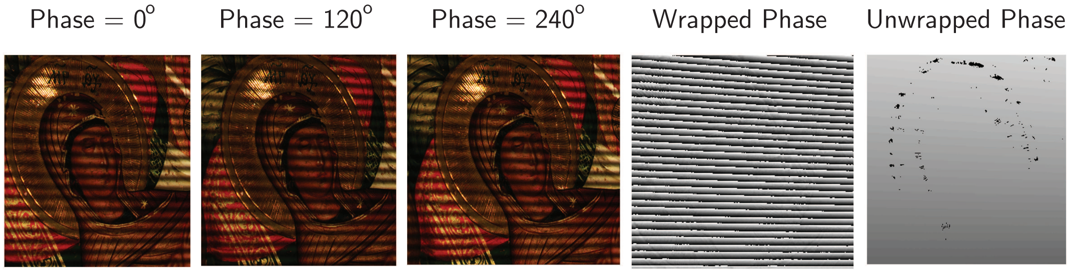


Figure 6-8: Sequential steps in fringe projection. In this case the three phase method has been used, after which three phases are wrapped with Equation 6-1 resulting into the wrapped phase image. Finally, the wrapped phase is then unwrapped.

6-4 Phase Unwrapping

After a certain amount of fringes are projected, the phases are wrapped using the formulas from Section 6-1-2. These wrapped phases have to be unwrapped in order to yield a usable continuous phase. A one dimensional example of this is depicted in Figure 6-1, and the implementation on an image in Figure 6-8. By definition, phase unwrapping is the reconstruction of a function on a grid (wrapped phase) bound by modulo 2π of the function on the grid (unwrapped phase). In ideal cases as displayed in the one dimensional illustration, phase unwrapping seems trivial. In practice, phase unwrapping is a challenge, and there are a multitude of methods to do this.

Small errors in the wrapped phase map can yield a sudden data point that is discontinuous, and a simple algorithm might wrap the phase at that point, leading to a large phase error that will cascade along the image. In the phase wrapping stage, parts of the images will be ignored when their signal-to-noise ratio κ (Equation 6-8) is too low. This might lead to disconnected phases, in which case our algorithms will have no idea of how the different independent parts of the phase image are related. Fringes that are occluded from the observer can be missed and mistaken for the next fringe (missing one entire wrap as discussed in Section 6-1-1). This in turn will result in a big and cascading phase error. Smart algorithms are able to 'wrap around' such occluded fringes, if these fringes can still be observed elsewhere. This is in fact often the case in paintings where big blobs of paint are local phenomena.

Smart unwrapping algorithms can cope with most of these problems, each with their own characteristics. There is no generic 'best' unwrapping algorithm. In testing different algorithms for our purpose, simple algorithms will inevitably fail because of the previously described problems that are abundant in the highly versatile surface of paintings. The assumption that we can make is that the surface we observe is continuous and connected, and no fringes are globally occluded. This makes an ideal case for 'guided path' [105] or network [106] algorithms, which choose an unwrapping path based on the probability that a wrap is a valid one. This probability is high when phase difference steps are small and the path is smooth and regular. This allows for partly occluded or dismissed parts of the fringe image to be accurately accounted for. The problem with 'smart' algorithms is that they are very slow, increasing exponentially with larger images. Implementations of this kind of algorithms are slow in languages like MATLAB, as it is not optimized for many calls to 'for' loops and 'if' statements. This can be overcome by implementing such a function in more efficient languages like C++.

6-5 Absolute Coordinate Mapping

Once the (unwrapped) phase image has been obtained, the carrier-related phase component has to be removed. This carrier makes the phase ever increasing, as is very apparent in the bottom graph in Figure 6-1. When the fringes are equally spaced, this carrier phase can be easily removed from a phase image through plane fitting (linear technique). However the fringes will not be equally spaced, because the projector and observer look at a common plane with a different angle, resulting in the need for a non-linear removal technique [107].

After removing the carrier-related phase, the shape-related phase is left and can be converted to real world coordinates. There are straightforward implementations available that assume that the camera and projector lie in a common plane parallel to the object's plane. In practice, this is hard to achieve and failure to do so will result in errors, especially when a high resolution is required. Such methods also involve the usage of a reference plane (that can also be used for removal of the carrier-related phase), but this is hard to implement in our particular system where large paintings are used and our cameras depth of field is very limited. The sole shape-related phase image can still be used to visualize relative differences in dimension.

6-6 Fringe Projection for Image Encoding

The unwrapped phase image has the same relative values per spatial location, irrespective of the viewpoint of the observer. In other words, any two real-world points that lie in the field of view of both the projector and observer, will have the exact same phase difference. When observing a scene from different viewpoints, their absolute phase value as computed by the unwrapping algorithm will generally differ, because this is dependent on where the unwrapping algorithm chose to start unwrapping, since unwrapping is a relative operation. Because there can be a large amount of fringes projected on the surface, distinguishing each separate fringe irrespective of viewpoint is non trivial. Moreover, when these are distinguished, one still has make out where the exact starting point of such a sinusoid is in order to obtain an absolute phase value.

Only when we have two values in (two) images with a different viewpoint that we know that corresponds, we can scale the phase values in both images accordingly, so that each phase value point in one image corresponds to the other. Such corresponding encoding is native in structured light methods like binary or gray-code projection methods, where each spatial location is labeled in an absolute fashion. However, those techniques other than fringe projection require a large amount of projections and are dependent on the projector's resolution as discussed in Section 4-3-5.

When the phase values in one image for one spatial location are the same for another viewpoint, we can use fringe projection as a scene encoder, where each observed spatial location is encoded with a unique value, irrespective of viewpoint. Two of such viewpoints (cameras in stereo) will allow for dense matching of such points and their triangulation, yielding its 3-D point in space. This type of encoding then solves the correspondence problem in stereo matching with a high accuracy and speed. Note that applying this method for two dimensions (as in an image), we need to encode the values in two dimensions as well if we wish to uniquely label each pixel. In practice, this means we need to project both vertical and horizontal fringes. Another important feature of using this method is that we do not need to characterize the projector's parameters anymore, as the observed luminance of the surface is the same irrespective of the observer's viewpoint (given that the specular reflections are supressed). This holds true for both the internal and external calibration parameters. Gamma correction does not have to be accounted for either, as the error resulting from this is the same for each viewpoint. As stated in Section 6-1-3, the projector's resolution is of lesser importance in fringe projection. Therefore, the choice in projector for this purpose is very liberal, even more so because we do not necessarily need to characterize it when using fringe projection as image encoder.

6-7 Conclusion

The classic method of fringe projection can yield depth data with a very high accuracy. Fringe projection is independent from the projector's resolution, so that we can capture a large amount of data if a camera with a large sensor is used. However, the classic implementation of fringe projection poses many problems with its performance as it is heavily dependent on an excellent projector characterization (calibration and gamma correction) and processing steps (coordinate mapping and carrier frequency removal). Using the classic method with two cameras (stereo fringe projection) will obviously yield another viewpoint, but the resulting errors will be present in both views. However, fringe projection can also be used as an encoding method, where each real world location in the view of both the projector and observer is uniquely labeled by a horizontal and vertical phase value. If we now introduce another camera in stereo, we can use this phase encoding to solve the correspondence problem. Doing so removes the need for exact projection characterization altogether, and the need for removing the carrier frequency component, as these are independent from the cameras viewpoint. The problem with this method is that this labeled phase value will not be the same in an absolute fashion, dependent on where and with what value the unwrapping algorithm starts the unwrapping process. This means that the phase difference value between any two spatial locations from any viewpoint is the same, although their absolute value might not be. In order to make this encoding method of use, these first have to be scaled such that the latter is true, and each phase value in one image corresponds to the same point as observed by the other image having the same phase value. In Chapter 7 we investigate the fusion of encoded fringe projection and stereo vision, in order to alleviate the shortcomings of both methods.

Chapter 7

Hybrid Depth Perception

We have seen in Chapter 5 that we can efficiently solve the correspondence problem in depth acquisition from stereo cameras through actively encoding the scene. We can encode the scene by projecting a structured light pattern onto it as discussed in Chapter 6. Within the structured light classes, fringe projection is an ideal candidate for our purpose because it constrains the total system the least.

The fact that the resolution of the projector poses no constraints to the fringe projection method is particularly important in a hybrid system where both cameras and projectors are used. Where the amount of pixels on camera sensors easily runs into the magnitude of dozens of megapixels, high-end digital projectors are in the range of two megapixels for a '2K' projector. This fact alone would justify the usage of fringe projection as the preferred structured light method when using it together with multi-megapixel cameras.

Since we wish to capture a high resolution over a large area, such multi-megapixel cameras are preferred as we should avoid data stitching with multiple captures as much as possible. Alternatively, objects that span a small area would not have to be stitched at all when sufficiently large camera sensors are used.

Due to its repetitive (ambiguous) nature, fringe projection is not suited to encode a scene in the way we require it. As opposed to many other structured light methods that encode a scene in an absolute fashion, fringe projection does this more relatively. The computed phase values that are produced by fringe projection generally differ slightly when the viewpoint is changed. In order to use fringe projection as an image encoding method, this has to be accounted for.

7-1 Fringe Projection Aided Stereo Matching

The usage of fringe projection for solving the correspondence problem has been suggested earlier [68] and tested in comparison with a binary structured light encoding technique. The possibility of acquiring a higher resolution using fringe projection was not explored, and the

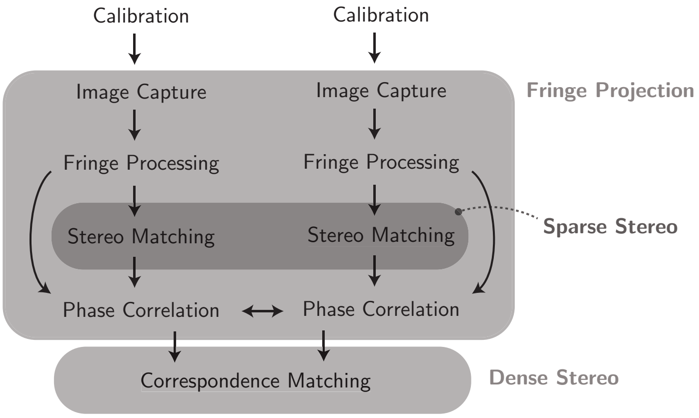


Figure 7-1: Schematic of the proposed hybrid combined depth and color acquisition method.

.

results pointed out that even for a similar resolution, the fringe projection accuracy was worse and suffered from artifacts. Moreover, in order to solve the absolute encoding problem in fringe projection, they employed the projection of multiple frequencies, leading to a total amount of 100 projections for a single image. In theory, fringe projection should give us the ability to obtain an accuracy a magnitude better and faster than using binary projection as encoding method.

Because of the high potential of fringe projection, we provide a framework in sequential steps on how the methods of stereo vision and fringe projection can be efficiently fused together. The schematic of the proposed hybrid method is depicted in Figure 7-1 and each part will be discussed in the following sections.

7-1-1 Setup and Calibration

For stereo vision, we will need to accurately obtain the internal and external parameters of the camera setup, while for fringe projection we need that for both the cameras and the projector. The need for exact spatial parameters of the projector is not required for our usage of image encoding. However, including the projector in the calibration takes little effort and gives better insight into the setup's configuration. Therefore, we only need to calibrate this setup in a stereo camera and projector setup as shown in Figure 6-2.

The influence of the relative locations and orientations of the cameras and projector in this setup is a combination of the constraints of both methods. Ideally, the cameras should be located outside the path of the stray reflected light emitted by the projector or it will induce veiling glare. We can solve this by moving back the cameras or increasing the angle between the camera and projector. In stereo vision, a larger angle between the camera and projector (and the cameras both oriented towards the same point on the plane) implies that the cameras are spaced more a part, which means we can resolve smaller depth deviations with a higher accuracy. When this distance becomes too large, the out-of-plane resolution and accuracy (z) will be high, but the in-plane (x and y) resolution and accuracy will decrease. Further constraints will be induced by the features of the chosen equipment like lenses and sensors. These specifics and practicalities will be discussed in Chapter 9.

7-1-2 Image Capture

We have seen in Section 6-1-2 that the common fringe projection implementation only needs to project six fringes; three vertical and three horizontal (shifted 120 degrees per step). We incorporate the double-step method, where we will effectively project another six fringes, but phase shifted with half a normal phase shift (60 degrees). Apart from the advantages of the double-step method, we project the second step with a larger amplitude than the former. This means that for the second step our signal-to-noise ratio will be higher, but we will be able to capture a higher dynamic range. This is necessary for many paintings that exhibit highlights, and the modulation of the fringes means we need an even higher dynamic range.

7-1-3 Fringe Processing

After the capture routine, both cameras will have captured 12 images or 4 sets of 3 fringes. Per camera we need to process a fringe set. The output of such a fringe set are wrapped phase map φ, average intensity map I' and quality related maps κ and I''. Then the unwrapped phase map φ will be wrapped to yield the unwrapped phase map, only at locations where the quality maps have determined the data is valid. The average intensity map I' resembles a common image that will be used for sparse stereo matching. An example of the unwrapped and wrapped phase images of the same scene in both cameras is depicted in Figure 7-2. In this figure, notice the amount of noise in the wrapped phases that is then discarded in the unwrapped phase through the quality related maps.

Per camera we obtain 4 unwrapped phase maps, 2 horizontal and 2 vertical (due to the double-step implementation). These are then averaged to yield one horizontal and one vertical map, and these will be used to uniquely encode the image. However, the phase maps in both cameras are only correlated in a relative sense. In order to solve the correspondence problem, we need the phase images in both cameras to yield the same value for a given spatial point. We can do this by first finding spatial points that are visible in each camera and correlate to the same spatial point. This can be done through sparse stereo matching. After that, we can add or subtract a value in the phase maps of one camera to make sure the entire phase map is now correlated in an absolute sense.

7-1-4 Sparse Stereo Matching for Phase Correlation

We will find spatial points that are distinct and visible to both cameras through sparse stereo matching. For this we use the SIFT algorithm to describe key points in the average intensity map I' of both images. These feature descriptors are then matched between the two views, and it is checked whether they fit into our correct homography. For example, a point in the upper region of the left image can never spatially correspond to a point in the lower region of the right camera in our setup. We are left with a multitude of points that correspond. For each correlated point within the unwrapped region, we determine the phase difference between the phase images of both cameras. This phase difference is then accordingly subtracted or added to one of the phase images. When this is done, we have encoded our image in an absolute fashion. Each phase value in one cameras image will match the same phase value for the same spatial location in the other camera.

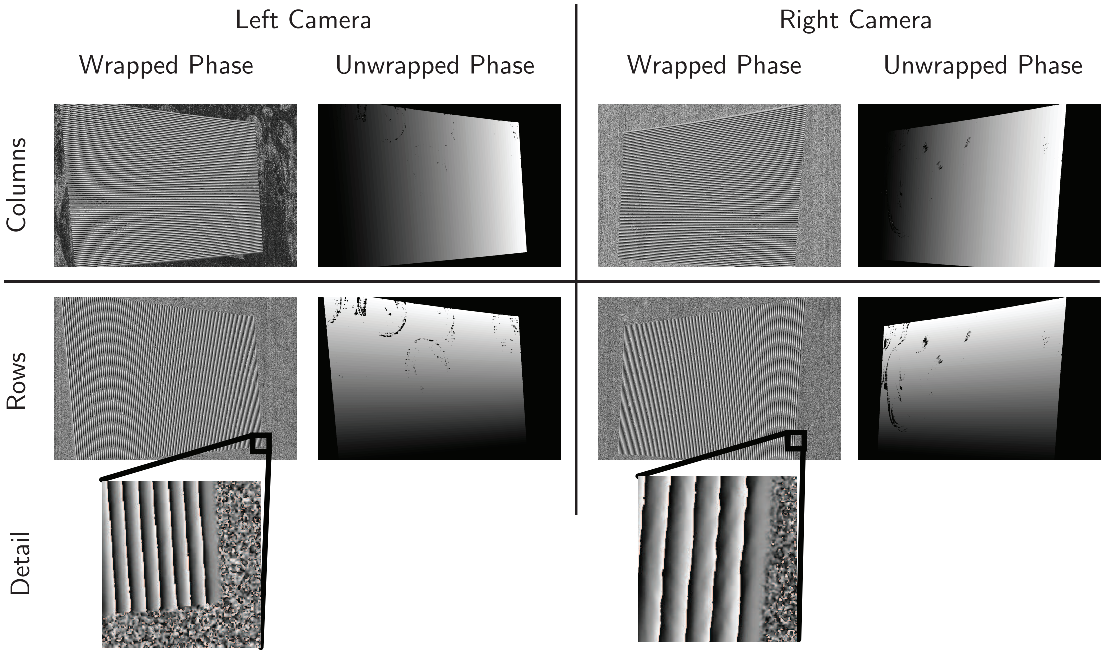


Figure 7-2: The wrapped and unwrapped phases for the left and right camera and the extracted phases (used for encoding) in two dimensions (rows and columns).

7-1-5 Correspondence Matching

Because the correspondence values are known, solving the correspondence problem in a dense stereo approach is now trivial. As we now have an appropriately encoded map per camera for the vertical and horizontal direction, we can proceed to match all the points that correspond, and triangulate those to yield a 3-D point. Note that each encoded point still has a pixel position, that corresponds to a texture color in the I' map. The problem is that our maps are still discretized, and spatial points generally are not captured at exactly the same pixel location, yielding a slightly different phase value. This means that we have to set a threshold on when two tentative phase matches should or should not be matched to one another. One such result is displayed in Figure 7-3. Finally, we will obtain a 3-D point cloud of our captured plane. The orientation of this plane can then be estimated, and rotated onto our coordinate system after which we can mesh the in-plane dimensions x and y and bin the depth values z inside of them. This will then produce a depth and an RGB map.

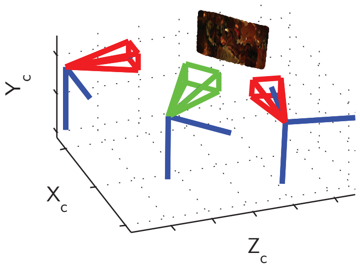


Figure 7-3: Each triangulated point can be visualized inside of the setup. In this way the validity of the triangulated points and the geometry of the setup can be easily verified. The physical setup corresponding to this particular scan is depicted in Figure 5-6(e)

.

Chapter 8

Image Registration and Fusion

In cameras, spatial resolution is inversely proportional to the size of the surface that is captured. The requirement of capturing a high resolution over a large area means that we inevitably have to make multiple registrations over this area. Image registration is the process of transforming multiple sets of data of the same scene into one coordinate system. If we encounter limitations posed by our hardware outside of the spatial domain, like the requirement of a large depth-of-field or dynamic range, we can use image fusion to enhance the constraints of our system.

8-1 Image Fusion

Image capture can be enhanced through fusion of multiple images of the same scene, by changing variables in the setup. Image fusion is the process of combining data from two or more images (with different settings) of the same scene into a single image. This can for instance lead to super-resolution[108], dynamic range (HDR) or extended depth of field (hyperfocus)[109]. The latter is often employed in macro-scale systems like microscopy where the depth of field is very small, and the environment is automated and rigid. Image fusion naturally requires one to capture multiple images, and process them. Excellent image registration is mandatory in order to provide good image fusion. Even with a good registration, whilst image fusion might enhance the overall average image characteristics, often it can also locally degrade well-registered parts in single images[110]. The previously discussed image fusion methods can be used for objects where an extension of the limitations of the camera or projector is required, although these might in turn pose different problems.

8-2 Registration and Stitching

The data that is captured through our method is in the form of a 3-D point cloud, although the data can be seen as 2.5-D since we are working with a plane with depth deviations. This

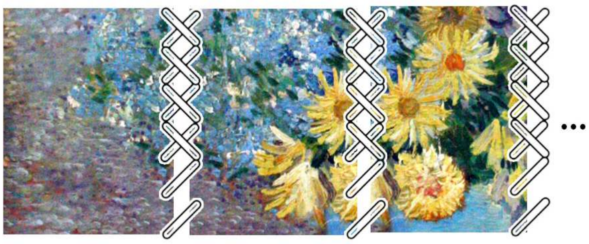


Figure 8-1: Illustration of the principle of image stitching. Multiple images are merged together to form a larger one. It should be noted that stitching of image maps can obviously be done in more than one direction.

2.5-D data can therefore be meshed into a 2-D map, yielding a depth map and a common RGB image map. If we wish to stitch multiple images together, we can choose to either stitch these maps, or stitch the 3-D point clouds.

One important remark is that we have a very good initial estimate of the location of the different captures, because the environment is static with respect to the observer. Each movement of the system is decided upon by the operator and closely monitored. Since the captured point clouds are calibrated 3-D points with the same units (mm), the operations we have to do on the point clouds will be rigid (translations and rotations). This, and the fact that we have a good initial estimate of its location reduces the registration complexity.

8-2-1 Image Data

Before stitching, we can choose to convert each 3-D point cloud to a 2-D depth map and color image map. The reduction of one dimension reduces the complexity, and moves the technique into a more common stitching domain (that of images as opposed to point clouds).

We give an overview of the methods we can use to stitch such images together. The process of image mosaicing where one constructs a larger image from a series of partially overlapping images is called image stitching. Image registration is the process of overlaying multiple images of the same scene. For our purposes, this method is probably inevitable due to the desired high resolution and large surface that may exceed an image equivalent of 100 megapixels[1], more than any commercially available RGB sensor can capture, let alone a depth camera.

Note that in stitching the depth map, terms like ‘exposure’ and ’gradient differences’ mean differences in the geometry and orientation between camera and scene, instead of brightness exposure due to lighting conditions or vignetting with an RGB camera.

The most common problems with the process of image registration is the intensity shift and geometric deformations between adjacent images[111]. Approaches towards solving the image registration problem can be classified according to their nature; area-based (dense) and feature-based (sparse). The majority of the methods consists of the following steps[112].

1. *Feature detection.* Salient and distinctive features are extracted from the image. These can encompass things like edges, points, lines, corners, etc. In an area based method,

[1] In the case of an 1 m^2 scene with a 300 dpi spatial in-plane resolution.

this step is omitted. Among the features are the single invariant descriptors like corners or points. The relaxation method labels each feature in one image to the other and iteratively repeats this process until it stabilizes.

2. *Feature matching.* When the features have been extracted they should be matched between multiple images with a good correspondence. In feature-based systems these can be matched by a specific matching function dependent on the type of features that are extracted.

   Area-based methods can be computationally expensive, but can be become faster through a course-to-fine matching strategy (pyramidal approach). Classic approaches are windowed correlation-like methods. Fourier methods are faster and resistant to frequency-dependent noise and illumination disturbances[112]. A problem with feature matching is feature ambiguity, yielding similar features. In our case, this can become a problem with repetitive craquelure. Among methods that use spatial relations are clustering methods, which match points by 'local evidence' that is connected by edges or line segments[113].

3. *Transform model estimation.* Estimation and construction of the 'mapping functions' are done on the basis of corresponding points. In order to reduce the complexity of this mapping function, it can be advantageous to correctly calibrate the camera. This can reduce complexity by using a simple mapping function instead of complex elastic and nonrigid transformations[114].

4. *Image resampling and transformation.* The previously acquired geometric functions are applied to the image. This transformation mapping itself and the application of it are often similar between different methods and are of lesser impact and importance than the first two steps of registration.

Another technique that might be applicable in our system is the area-based mutual information (MI) method. Such a method is suitable for the registration of images from different sources, like matching a depth map from one source and an RGB map from another. In such a case one might make a rendering of the depth map using artificial lighting, and match features like points, edges and corners[115]. Due to the different nature of the image and the problems involved, such matching is best avoided if possible.

Area-based methods in general only allow for smaller shift and rotation between images, so a good estimate of their overlap would be preferred. Feature-based methods usually allow for large rotations and displacements, and can handle complex distortions. The feature-based methods are highly dependent on distinct and unambiguous features.

After the geometric image registration has taken place, the images have to be fused to yield an image with preferably no obvious seams and artifacts that might challenge the illusion of homogeneity. This is the biggest limitation of the registration of images as opposed to the point clouds. A rotation around the z (depth) axis is not a problem, as this is solved by a simple rotation of the 2-D map. Rotations around the x and y axis however, will result in difference in depth over the image, consequently resulting in a gradient in the depth map. Since the amount of depth deviations we have is very small, even a very small rotation of the setup with respect to a previous image will induce such gradients. These gradients are hard to filter out using 2-D image stitching methods, especially if we want to stitch multiple captures made over two dimensions (when the setup is translated in width and height).

8-3 Point Cloud Data

Since we have multiple captures with a good initial estimate and with similar units, the only operations we have to make on the different data sets are a translation and possibly a small rotation. This means we can limit ourselves to rigid transformations, reducing the need for complex algorithms and increasing the reliability of its solution. The increased dimension as opposed to the image stitching method does increase complexity, especially given the large amounts of data we capture per image (a magnitude of millions per capture for the proposed hybrid method). However, the usage of an arbitrary implementation of the point cloud registration method might not suffice, as in our case we wish to stitch planes with a large width and height (magnitude of decimetres), but a relatively minimal depth deviation (magnitude of millimetres). Stitching the planes themselves of the multiple point clouds should be easy for any method, but making sure the small depth deviations overlap may require us to adapt the method in terms of setting a very low solution tolerance.

A straightforward implementation that works well given a good initial estimate and the ability to converge to an optimal solution is the iterative closest point (ICP). The basic implementation of this algorithm uses rigid transformations (translations and rotations). It works through first associating closest points, estimating the best transformation parameters (using a certain cost function) and applying those, after which the process reiterates. A convergence graph towards an optimal solution over the iterations is given in Figure 8-2. As an initial registration step, we rougly align the two point clouds dependent on where our linear axis has been translated to. We can see this initial estimate is quite good, with only a root mean square (RMS) distance deviation of 1.5 mm for a 15 cm translation. The ICP algorithm then makes the solution converge to around 30 µm. The ICP example from which the graph originates is shown in Figure 8-3. The ICP process can be repeated for multiple adjacent point clouds to span an even bigger surface. This method iterates towards a local solution, dependent on the initial estimate. Therefore, the initial estimate has to be good enough, or it might converge to an erroneous location. As opposed to previously mentioned mutual information methods in image stitching, this algorithm does not assume our data points are each labeled with texture information (color). However, the algorithm is very adaptive and the cost function estimation can be adapted to include a color difference metric [116].

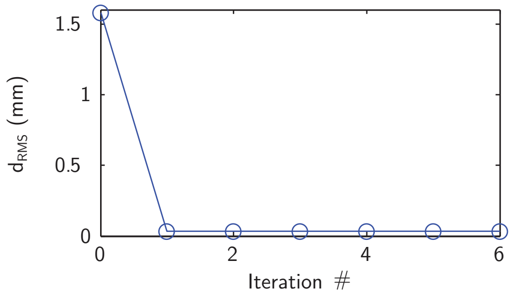
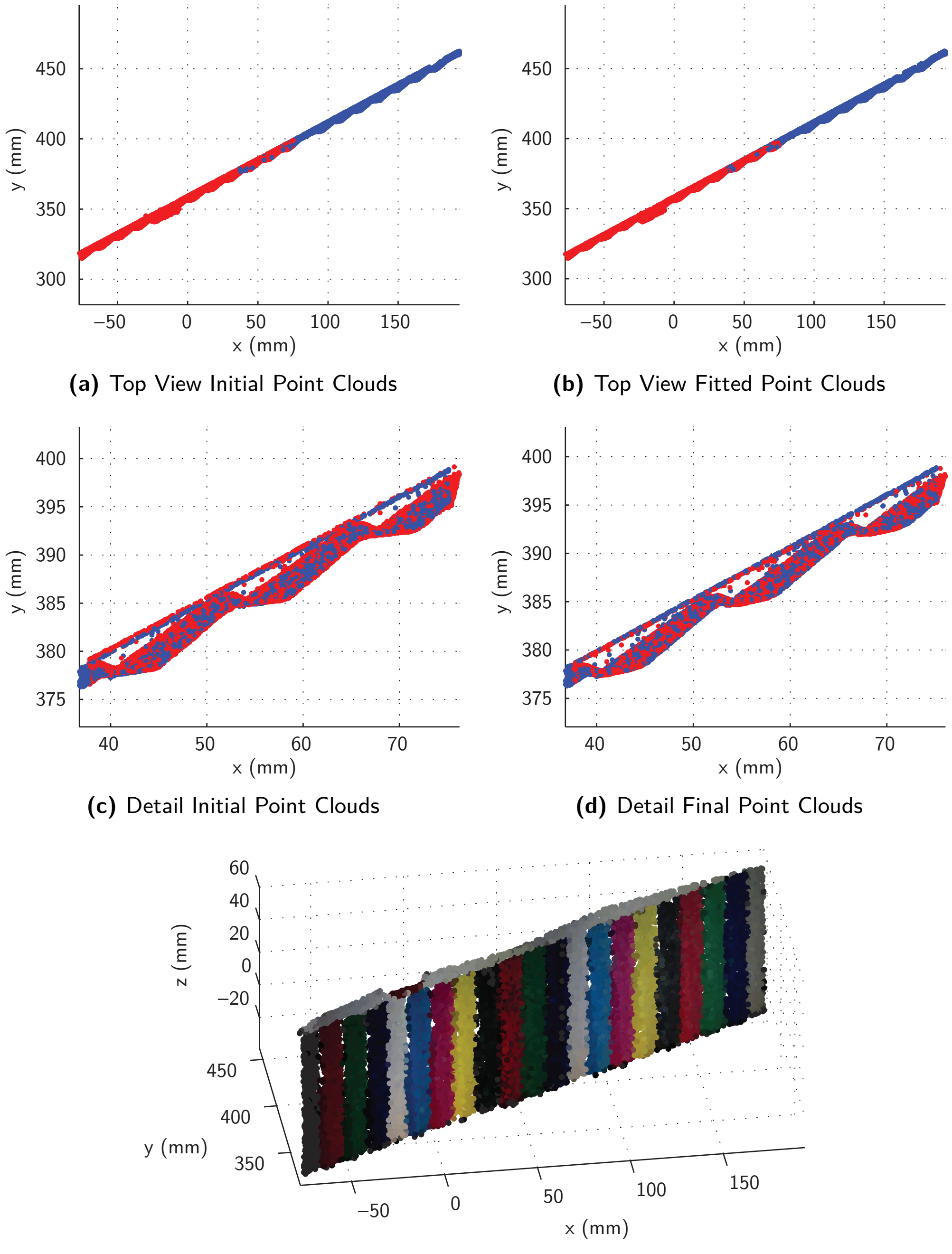


Figure 8-2: Convergence to a solution of aligning two point clouds using the ICP registration method.



(a) Top View Initial Point Clouds

(b) Top View Fitted Point Clouds

(c) Detail Initial Point Clouds

(d) Detail Final Point Clouds

(e) Combined Textured Point Cloud

Figure 8-3: A top view of two point clouds with a initial fitting estimate seen in (a) and a detail of their overlap in (c) are aligned using the ICP algorithm. We can see the result in (b) and a detail in (d). The final combined and textured point cloud is depicted in (e). Note that here vignetting is not accounted for, resulting in a darker appearance of the texture around the edges of both former separate captures. The 50 million points contained in this point cloud have been resampled to 10.000 for the above illustrations.

Chapter 9

Design Implementation

The final design of our system was an iterative process. Things that are stated in literature to work well, might indeed work in certain conditions or for certain hardware, but turn out to be impractical in terms of implementation in a system for our specific purpose. Inevitably, new problems arise that are specific for our hardware, implementation or purpose. These real-world practicalities and the choices and implementation arising from them are described in this chapter.

9-1 Hardware

Each element in the practical design has to be suited for our application and work together with the rest of the equipment in order to constrain our system the least. The constraints and limitations of our system are imposed by the equipment. Since we do not only wish to test and characterize the performance of our design, but also deploy it in practice, we need to use the best equipment available. Each component of our system is described in this section. A frontal view of the scanning head on a linear translation axis is depicted in Figure 9-1, and a top view with approximate fields of view in Figure 9-2.

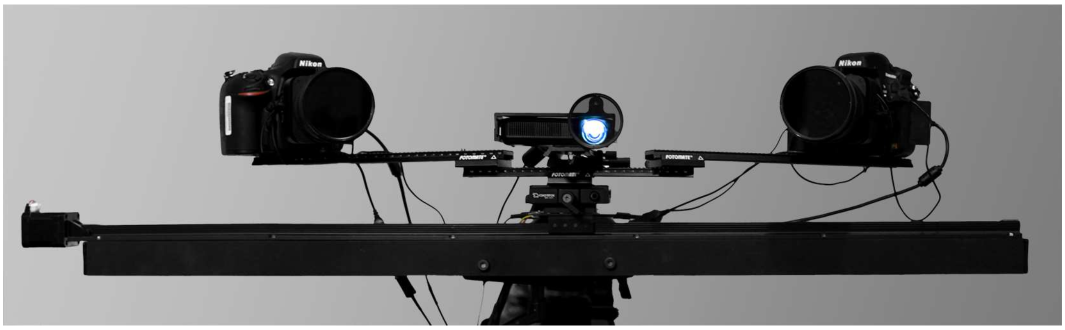


Figure 9-1: The imaging 'head' mounted on the linear (x) stage.

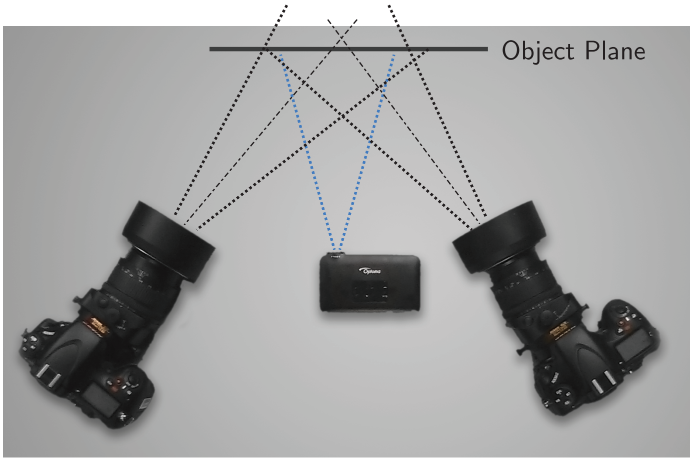


Figure 9-2: Top view with rough fields of view of each device. Notice that the special lenses are slanted to adhere to the Scheimpflug condition.

9-1-1 Spatial Movement

Because there is no sensor that can capture the large amount of data needed to capture a painting in its entirety in a high resolution, we need to translate the system. We need to move around the canvas in the width (x) and height (y) direction. Because our image stitching method (Chapter 8) is fairly dynamic, we do not have to translate with millimeter precision.

For this first implementation, we chose to only automate the x direction. The y direction can be done manually by mounting the system on a studio tripod. Automatic translation in this vertical direction would have significantly increased the structural complexity and cost. For the translation in the x axis, we chose a 110 cm long translation axis with a screw spindle that had a thread width of 2.54 mm. The axis was driven by a stepper motor with 200 steps per revolution, but our stepper driver allowed for 3200 steps per revolution in 'micro-stepping' mode. The axis we bought was actually a bit too small because we initially chose it to carry a laser profilometric scanner, which was very light, but its performance was not satisfactory.

9-1-2 Cameras

For image capture we needed color cameras with a good dynamic range that could be interfaced with and controlled by the computer. Because we would like to capture as much information as possible per shot, we preferred a large sensor with a high resolution. Cameras that ideally fit this description and are relatively cost effective are consumer Digital Single Lens Reflex (DSLR) cameras.

In our first implementation we used two NIKON D80 camera bodies with 50 mm lenses. These featured 12 megapixel sensors, and can be picked up second hand for around € 150. Tethered control of these cameras was possible (though cumbersome), and the images they produced were very satisfactory. All settings of the camera were set to manual and could be set by the computer.

Because the projector is in front of the object plane, and the cameras would be viewing this plane from an angle, the cameras need a very large depth of field, so the aperture value was set

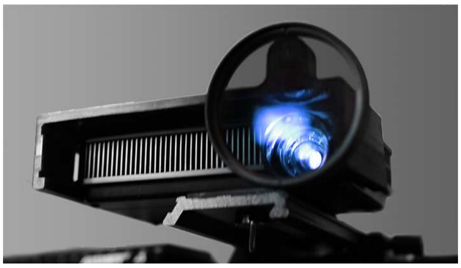


Figure 9-3: Polarization filter for the projector on a 3-D printed housing.

to be very high (around f/20). Such high aperture values are not optimal for image quality. Therefore, we managed to borrow one 85 mm PC-E ’tilt-shift’ (Scheimpflug) lens, and bought another. By coincidence we were able to borrow two NIKON D800E camera bodies, featuring 40 megapixel (full-frame) sensors. The retail price for one such camera-lens combination is around € 4000. Therefore we used the latter camera-lens combination.

9-1-3 Projector

Small (pico) projectors are ideally suited for our system, because they are small, feature a very short throw distance, and are cheap. The DLP projection technique was prefered as the light it emits is unpolarized (needed for our cross-polarization setup). The effective resolution of the projector is of less importance, because we will eventually slightly defocus the fringes. We therefore bought an OPTOMA PK301 projector second hand for € 200. The projector should be turned on a few minutes before operation since its performance might change under the influence of temperature changes.

9-1-4 Cross-Polarization

The lenses on the DSLR cameras had a convenient fitting to accommodate for polarization filters. These were oriented in such a way that surface reflections from the plane in front of them were optimally suppressed. We achieve cross-polarization by putting another polarization filter in front of the illuminant (projector), oriented so that it is polarized perpendicular to those of the camera. This will yield an even larger decrease in surface reflections, so effective that (at least to the eye) it eliminates reflections completely. As our projector did not have a convenient filter fitting, we 3-D printed one as shown in Figure 9-3.

9-1-5 Controller

In order to tie together all the hardware discussed above, we need a controller. The controller was fabricated as a Printed Circuit Board (PCB) because of its high reliability. The requirements and their solutions as implemented in the PCB for this controller are as follows.

- **Communication** The communication between the computer and the control board should be straightforward, simple and universal. The obvious choice is using the USB interface, as it provides two way communication and a reliable 5 V DC power supply.

- **Processor** The component in the PCB that ties the components together should be able to communicate with all of them and should be able to be easily programmed. The obvious choice here is an ATMEL chip flashed with the ARDUINO bootloader that allows programming in the C++ language. This can be tied to the USB port through an FTDI chip.

- **Stepper Driver** The translation spindle's stepper motor has to be driven by a circuit. This is best done through a separate specialized chip, for example the ALLERGO A4988 that allows for micro-stepping. The motor will need a separate 7 V DC power supply. The conversion of distance-to-steps is done by the processor, and we save this position in the static EEPROM memory, which persists when the system is turned off.

- **Manual Operation** The setup should be manually adjustable, and in order to move the translation axis, we have to control the stepper motor. This can be done through a simple joystick mounted on the PCB, communicating with the processor which in turn drives the stepper motor. At the far ends of the linear stage, magnets are mounted. These are sensed by two Hall sensors on the PCB, so that it knows when the translation limits are reached. These are also used to calibrate the absolute spatial position with respect to the translation axis itself.

- **Pertubations** The fact that the setup should be portable means we cannot use very heavy (and sturdy) constructions. Therefore, the possible lack of sturdiness should be accounted for. This is possible by measuring perturbations through for example MEMS devices as the MPU-6050, which features a combined 6-DOF accelerator and gyroscope. A high motion amplitude should postpone the capture of an image, as these perturbations lead to blurred images with longer exposure times.

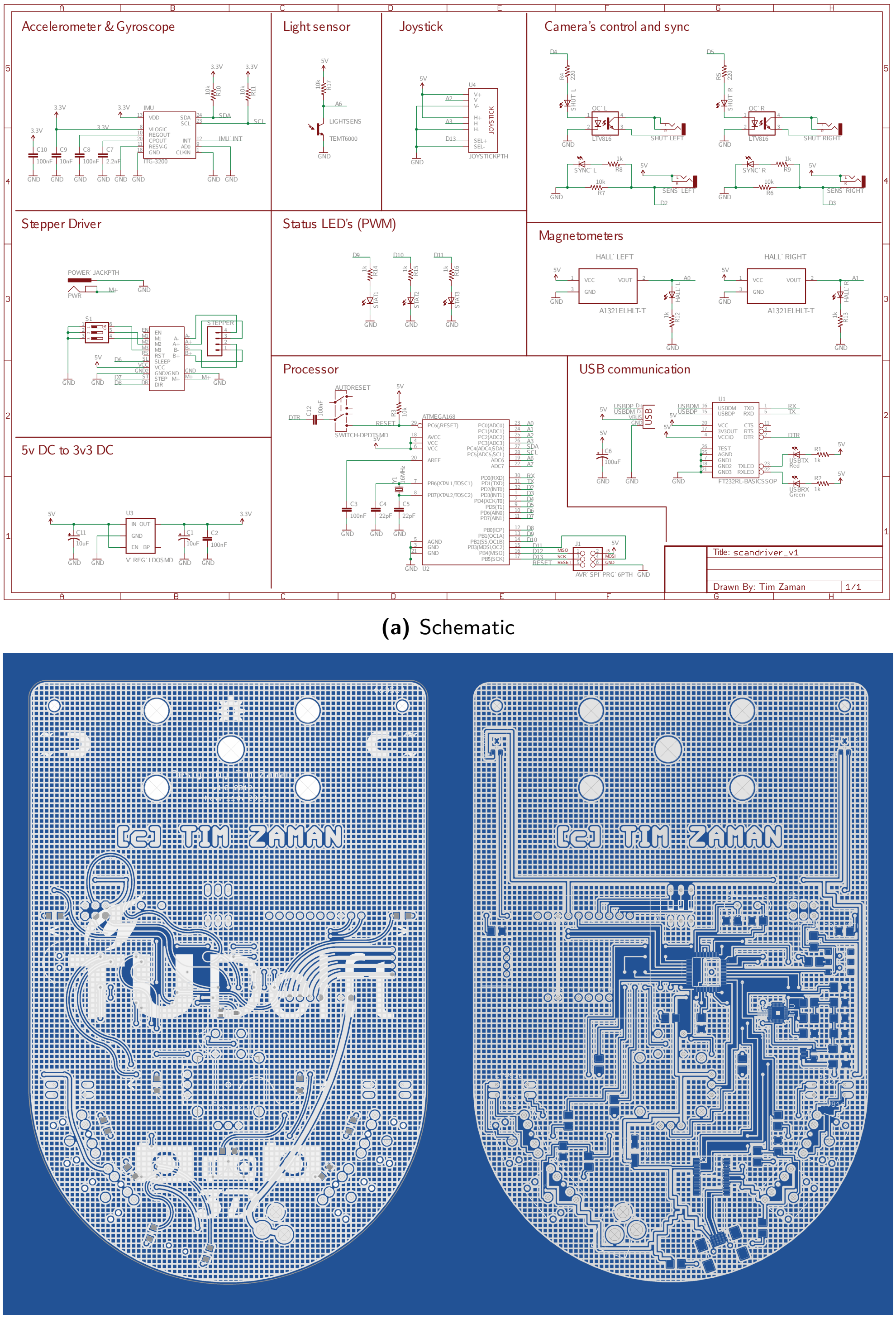
- **Cameras** The host computer should externally trigger the cameras through the PCB. This is best done through an optocoupler as it separates the cameras triggering circuit with that of the host. Whether the cameras have actually triggered can then be sensed through connecting a synchronization (or flash-) cable between the two devices.

These requirements and components eventually lead to the device's schematic and board depicted in Figure 9-4.

After careful revisions, the board was produced in Shenzhen, China by SEEEDSTUDIO. Most of the components are in the Surface Mount Device (SMD) form-factor to allow for a small footprint, and were soldered by the author using hot-air rework. The final populated PCB is shown in Figure 9-5.

9-1-6 Final System

The system's equipment as assembled is depicted in Figure 9-1 (although not mounted on the vertical tripod). Because there is a trade-off in captured area versus spatial resolution, we chose to go for maximum resolution first. In that way, we will obtain a higher resolution than strictly necessary following our requirements. This is because we do not have a feeling yet on how high the resolution of the final result, a 3-D print, has to be in order to appear convincing. If it finally turns out that a higher resolution is necessary than we have obtained, our captured images will be rendered obsolete. Inversely, high resolution images can more easily be down-sampled to a lower resolution. Therefore we positioned the cameras as close to the object plane as the lenses allow for (50 cm). These are then positioned symmetrically on both sides of the projector that both observe the plane with an angle of 20° with respect to the projector. Increasing this angle more will induce the chance of occlusions and yield an



(a) Schematic

(b) Board layout, double sided with the top and bottom on respectively the left and right.

Figure 9-4: The PCB design featuring the schematic in (a) and corresponding board in (b).

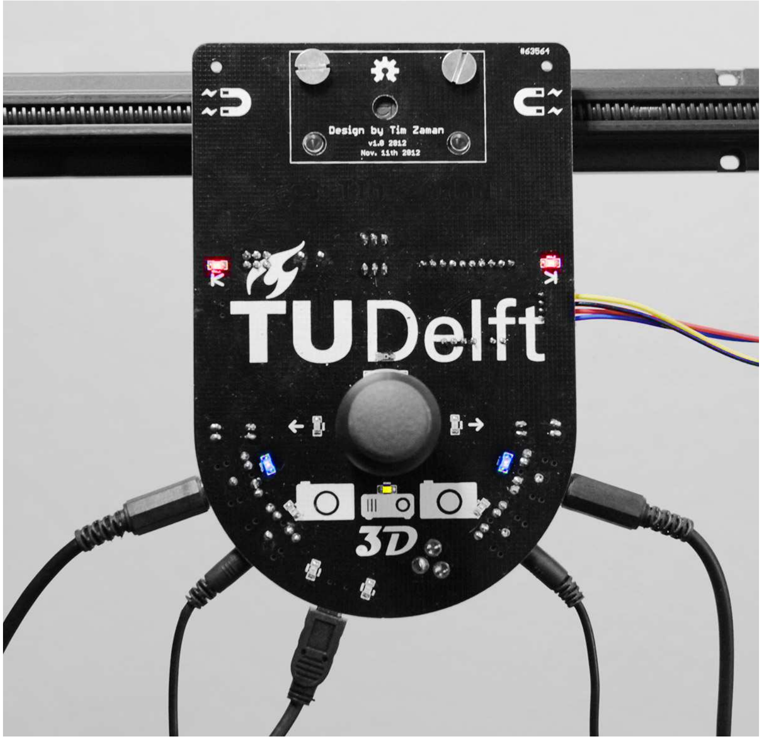


Figure 9-5: Top-view of the populated controller.

increase in depth accuracy z, but at the cost of a drop-off in the width accuracy x in relation to the height accuracy y. The projector is oriented perpendicular to our object plane, to fall inside the fields of view of both cameras. The exact spatial location of our setup might vary some depending on our object of interest and requirements, for which our highly dynamic calibration methods allow.

9-2 Software

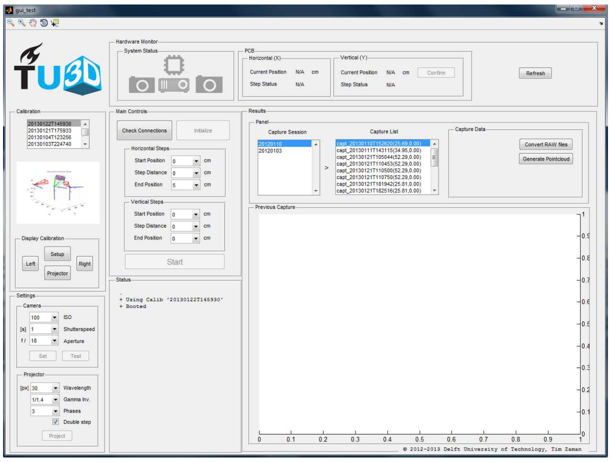
Most of the software was written in MATLAB for its ease of prototyping. Some functions were written in C++ and compiled to MEX functions that can be invoked from MATLAB.

9-2-1 Graphical User Interface

Because of the large amount of settings, functions and data in our system, a Graphical User Interface (GUI) was designed from which functionalities can be invoked, settings tuned, and data captured and visiualized. Such an interface is easily conceived using MATLAB. A screenshot of the GUI during start-up is displayed in Figure 9-6.

9-2-2 Calibration Routine

We characterize the color capture and spatial model of the entire setup through one single calibration routine. When this is done, we have internal models of the three optical systems



Horizontal (X): Current Position N/A cm; Step Status N/A

Vertical (Y): Current Position N/A cm; Step Status N/A; Confirm

Calibration:
20130122T145930
20130121T175933
20130104T123256
20130103T224740

Horizontal Steps: Start Position 0 cm; Step Distance 0 cm; End Position 5 cm

Vertical Steps: Start Position 0 cm; Step Distance 0 cm; End Position 0 cm

Camera: 100 ISO; [s] 1 Shutterspeed; f / 18 Aperture

Projector: [px] 30 Wavelength; 1/1.4 Gamma Inv.; 3 Phases; ☑ Double step

```
-
+ Using Calib '20130122T145930'
+ Booted
```

Capture Session:
20120110
20120103

Capture List:
capt_20130110T152620(25.69,0.00)
capt_20130111T143115(34.95,0.00)
capt_20130121T105044(52.29,0.00)
capt_20130121T110453(52.29,0.00)
capt_20130121T110500(52.29,0.00)
capt_20130121T110750(52.29,0.00)
capt_20130121T181942(25.81,0.00)
capt_20130121T182516(25.81,0.00)

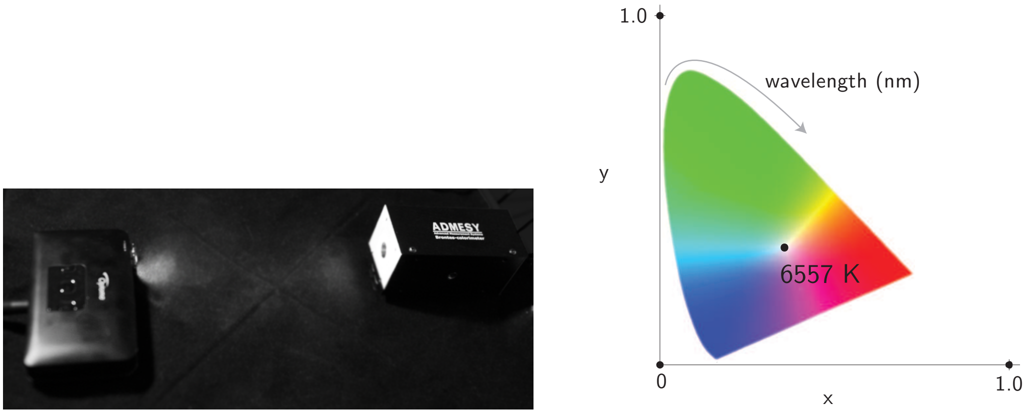
Figure 9-6: The graphical user interface 'TU3D' facilitates easy interaction with the hardware and routines.

(cameras and projector), and their relative spatial positioning and orientation. Furthermore, we generate an appropriate ICC color profile for both that relates their captured colors to the real-world intensities.

Color Calibration

For the color calibration, a characterized Colorchecker reference chart is presented to the system and captured and processed to an RGB map. We make sure the white and black patches fall inside our captured gamut (no clipping). From the GUI, the four patches in the corners are then selected and all 24 patches are automatically read out. The RGB values of these patches are then related to their reference LAB values and the projector's white point. The white point of the projector was measured to be around 6500 K (D65 standard illuminant) as illustrated in Figure 9-7, and is similar to the average color temperature during daytime in western Europe. These values are then converted by ArgyllCMS into an ICC color profile, relating our captured colors to our real world reference. This ICC profile will then be applied to each captured image through ImageMagick, and then converted to an internationally standardized ICC profile (eciRGBv2).

Lens limitations often lead to the reduction in brightness at the edges of the captured or projected image's edges when compared to its center. This phenomenon is called vignetting. Although some cameras can account for this fact internally, it is a good idea to model this behavior. The vignetting of the cameras can easily be modeled and accounted for by taking a defocused image of a spectrally neutral white balance chart under even illumination. For the projector we cannot repeat the same process, because we need to know the orientation of the neutral chart in order to trace back the projector's light rays. This can be done by capturing the white chart in 3-D using our hybrid method. We then obtain an image where each captured point has an observed brightness and can be traced back into the projector. A deviation of each of the brightness values in relation to each other hints to vignetting. A vignetting map can then be made. This map can then act as a scalar on the final RGB images.



(a) Measurement Setup

(b) CIE Chromaticity Diagram

Figure 9-7: Characterization of the projector as illuminant with the setup in (a) showing the projector (left) and colorimeter (right). The extracted color temperature in Kelvin is displayed inside a chromaticity diagram in (b).

Spatial Calibration

For spatial calibration, we use images of blue checkerboards with known dimensions that come straight out of the camera. The process of combined camera-projector calibration has been covered in Secion 6-2-2. Each checkerboard is simultaneously captured by both cameras, while the projector projects its own red checkerboard pattern on the blue checkerboard. Around 20 images are captured of the checkerboard in different orientations. In the GUI, we will now have to select the corners of each checkerboard. Note that this has to be done for both the real world blue checkerboard as well as the projected red checkerboard. The core engine behind the MATLAB Camera Calibration Toolbox was used for processing the feature locations of the checkered features and the calibration that follows. This toolbox was adapted to allow for projector calibration by inversion of the camera model (points-to-rays instead of rays-to-points). The estimation of the internal and external parameters is then optimized by the toolbox. Finally we proceed to pre-compute the optical rays for each pixel in both the camera's images. This means that each pixel captured in either camera has its own vector in space in relation to the global coordinate system, originating from the optical center of the device. This will allow for easy ray intersection between rays originating from both cameras later in the software.

9-2-3 Capture Routine

As with most of the previous third party applications, a wrapper had to be written in order to control and interface with the cameras inside of the MATLAB environment. The communication layer was made using GPHOTO2, that works with most DSLR camera models. The cameras are connected through a USB interface. In order to talk to multiple cameras that have the same USB identifier, LIBUSB was used to make the connection with GPHOTO2. The GPHOTO2 wrapper was written using the Java application programming interface (API) that is native to MATLAB. This all had to work together with the PCB that controls the other hardware in the setup. An image is taken by requesting a synchronized shot through a serial interface between MATLAB and the PCB. The PCB then simultaneously triggers both the camera's shutters. The projection of the fringes is done through hooking up the projector as a second display. This second display is then controlled from a toolbox called PSYCHTOOLBOX. This toolbox provides MATLAB wrappers for low-level OPENGL commands that can control the display. After one scene has been captured, MATLAB issues the command to the PCB for the translation axis to move to the next location. Camera, projector and fringe parameters can all be set from the GUI, as well as automated multiple captures at different locations along the x axis (the y axis is manual). When the y axis is translated, the operator measures the translation distance using an accurate distance measurement device (LEICA Disto D2).

The image processing that converts the images from their raw format to their final point clouds and maps is separated from the capture routine. The GUI has the ability to process all the captures afterwards in batch mode.

9-2-4 Data Processing

The schematic steps involved in our method are depicted in Figure 7-1. The same schematic displaying the steps in data processing terms and its functions is illustrated in Figure 9-8. The steps leading up to the dense stereo matching process have been previously discussed. The dense stereo matching was in need for a specific implementation, as the task of matching approximately 30 million points between two images forms a large bottleneck in our system. Each capture routine results in two absolute phase maps per camera; a horizontal and vertical one. These maps have been correlated to each other between the two cameras, so that each spatial point that is visible by both cameras has the same phase value. The problem is that these phase values are not necessarily identical, which makes the matching of the two nontrivial.

An implementation where each absolute phase label in one image is matched with the closest label (Euclidean distance) in the other takes 10 days of computation time per capture. The order of this algorithm is around $O(n^2)$, where n is the amount of pixels per image. Therefore, we first identically scale and round off the values in both the absolute phase maps such that each pixel approximately corresponds to a phase increment of 1.0. Our map is still represented in image coordinates, where each pixel location on the sensor (x, y) corresponds to an integer absolute phase value (h, v) (horizontal and vertical respectively). For correspondence matching, we are looking for the values (h, v) in the left image, that are closest or identical to (h, v) in the right image. When such a match is found, we use their respective pixel locations (x, y) in both images. Since the values of (h, v) are integers, we can invert the coordinates of our map, creating a look-up table, such that the dimensions are in terms of (h, v), and each location (h, v) in our map will yield a value (x, y). Doing so for both cameras means that every location in their look-up table maps corresponds to the (x, y) position in their sensors. In this way, we avoid using 'for-loops' and distance computations, although we have lost some precision in the discretization steps.

We can now construct a vector where the rows span the amount of matches n, and the columns indicate pixel locations $(x, y)_{left}$ and $(x, y)_{right}$. Then we substitute each pixel location (x, y) by their corresponding optical ray vector from the optical ray map we pre-computed in the calibration routine. This optical ray vector is expressed in the system's coordinate system,

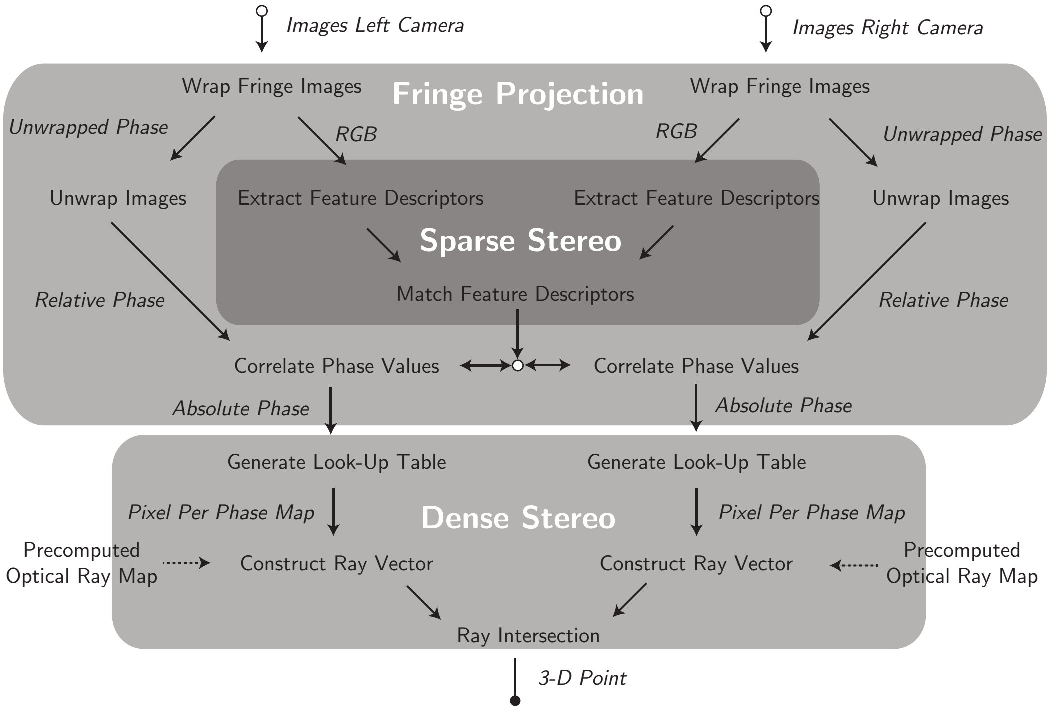


Figure 9-8: Schematic of the data flow for the hybrid method depicting the fringe projection fused with the sparse and dense matching. Annotated arrows in italics indicate the type of data that is moved while the normal font style indicates a separate function or process.

.

and originates from the optical center of each respective camera. The size i of our vectors is around 30 million for a typical capture, and both vectors can now be used for triangulation through ray intersection in a single operation. Because in practice the two lines (rays) in 3-D rarely exactly intersect, we will obtain a point for each line that minimizes the distance to the other line. A small distance will validate the triangulation, and the average location of both points will yield the single 3-D point per match. Matching the phase values in this manner leads to their triangulated points is in the order of $O(n)$. Comparing this to the regular closest point estimation loop of order $O(n^2)$, the factor of speed increase using the faster algorithm is equal to n. Since n can be as large as 30 million, we indeed find out that our algorithm takes seconds instead of 10 days. However, a prerequisite of the fast method is that a look-up table is constructed, which takes around one minute.

The output of our process is a 3-D point cloud where each point is labeled with a color RGB. For easy data validation, we estimate the plane orientation and mesh all the values into two 2-D (x, y) maps; a depth map (z) and RGB map.

Chapter 10

Results

In order to assess whether our design meets our requirements and digitization guidelines, we have to assess all aspects of its performance. For most aspects this is done through several reference objects. The data we provide to the 3-D printer will be in the format of a 2-D depth and an RGB map. All our tests will therefore be performed on such maps, as it is the final product of our system. The results that are published in this chapter are highly dependent on the configuration of our setup. The configuration from which these results were obtained were those that were used in practice in the scanning of a painting by Rembrandt. A setup was chosen where a maximum detail can be obtained by moving the system as close as possible to the painting, constrained by the minimum focal distance of the lenses. The configuration of the setup setup is illustrated in Figure 10-1.

10-1 Invasivity and Portability

The projector is the only component of the system that actively projects light on the object. This means that the projector is the only piece of equipment that is viable for violating our

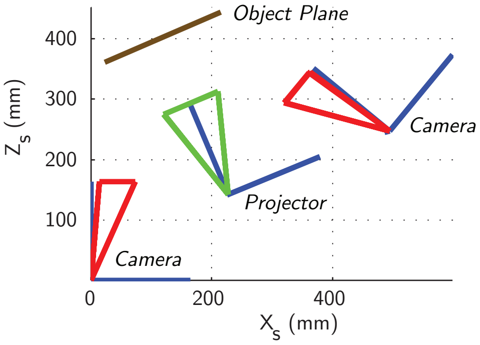


Figure 10-1: Top view of the setup's configuration as used in practice. The size of the viewing pyramids have been exaggerated.

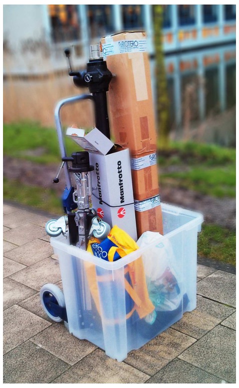


Figure 10-2: Trolley containing the entire setup, demonstrating its portability.

invasivity constraint. Our Optoma PK301 projector is stated to have a maximum brightness of 50 lumens. An illuminated area of 17 cm by 10 cm equates to a surface of 170 cm^2. The luminous flux per unit area is then $(50/170) \cdot 10^4 = 2941$ lux (normal office lighting is around 500 lux, direct sunlight around 100,000). This is vastly more than the generally accepted maximum luminous flux for sensitive museum items of 50 lux. However, this recommendation is made for continuous museum illumination of the objects, whereas one capture at 2941 lux will not take more than 2.5 minutes. We will not use the maximum brightness setting of the projector, and the amplitude of the fringes we project varies so the effective amount of lux will be less. More importantly, the time the light is emitted is in a trade-off with that of the flux [83]. This means that a large flux is allowed given that the application time is small. In that sense, our illuminant will be a factor more invasive than the normal illumination conditions, but would be considered non-invasive. Because the projector uses LEDs as illumination source, it should not emit a significant amount of UV or IR radiation. An IR laser thermometer indicated that the surface does not heat up with any significance, nor can heat be felt to the touch under this illuminant, agreeing with the guidelines in [83].

The entire system is modular, this means that it can be moved using a simple hand-held trolley as depicted in Figure 10-2. The electronic equipment can be carried separately in a normal sized backpack. Taking the setup apart takes half an hour, setting it up around one hour, followed by half an hour of semi-automated calibration.

10-2 Color Reproduction

Because a topographic imaging device is developed for the purpose of making accurate 3-D reproductions, the simultaneous capture of the surface color is a very important feature. The capture of both the depth map and RGB map means that these do not have to be separately matched, aligned, stitched or registered on top of each other, since every 3-D point is labeled with its own color. Therefore we assess whether this color information is accurate and precise enough to serve as texture information. This is done through an X-Rite Colorchecker chart with a known spectral response and reference values in the LAB color space. We will also assess the color channel misregistration using a QA62 slanted edge target.

For assessing the color performance, we capture both the reference targets as shown in Figures 10-3 and 10-5. These are then processed and meshed into a depth and RGB map (this is

Table 10-1: Individual color accuracy results (in ΔE-2000) per patch on the color chart.

	A	B	C	D	E	F
1	1.0	2.3	1.6	1.8	2.3	2.0
2	1.6	1.5	2.2	1.5	2.1	2.1
3	1.2	1.9	2.0	2.0	2.0	2.6
4	1.6	2.5	2.5	2.2	1.9	1.1

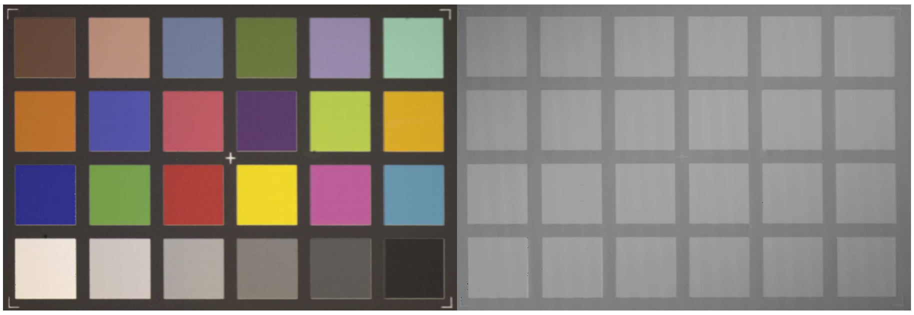


Figure 10-3: Reconstructed RGB map (left) and depth map (right).

also the way the data is presented to the 3-D printer). The RGB images are then converted to the LAB color space, taking into account the corresponding ICC profile.

In the case of the color checker, this is related to its reference values in the same color space using the same color temperature and white point (6500 K or D65 standard illuminant). For each patch, we obtain a mean value of the patch, and a standard deviation within the measured window (noise). We use the mean value to assess accuracy and the noise to determine its precision. The noise was measured in brightness channel L*, and had an average 8-bit deviation of 1.0 L*, or 0.39% over all the patches. This can be used as a metric for color precision.

The deviation between the measured and reference LAB values for each patch is given in ΔE-2000, a metric for color reproduction accuracy. The mean ΔE-2000 value over all the patches was 1.9 ($\sigma = 0.44$). The ΔE-2000 values per patch are given in Table 10-1, corresponding to the rows and columns of the chart in Figure 10-3. The guidelines for the digitization of cultural heritage (FADGI [82]) state that the mean ΔE-2000 values should be below 3.0, and the maximum per color patch 6.0 in order to receive the best quality rating. Our color reproductive abilities are therefore more than sufficient. It should be noted that this is only weighed over the 24 colored patches, like the guideline states. Reference charts that have more patches that cover more of the color gamut would give us a more accurate result.

Color channel misregistration as depicted in Figure 10-4 is measured using a QA-62 target, and is a phenomenon that reveals itself in color, but is a spatial feature. It is eiter induced by the demosaicing of the bayer color pattern in front of the camera's chip, or chromatic aberration induced by the optics. The average color channel misregistration we measured was on average 0.04 pixels as given in Table 10-2. The FADGI guidelines state that this value should not exceed 0.3. Our obtained value is very low because we used state of the art macro lenses that do not exhibit much chromatic aberration, although this value will increase somewhat when this is measured at the extreme edges of the image where it is generally more

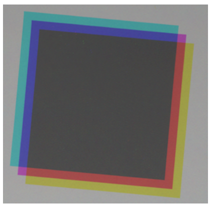


Figure 10-4: Exaggerated example of a grey square that exhibits color channel misregistration (red, green and blue channels have been shifted disproportionally).

prominent.

10-3 Spatial Reproduction

The spatial resolution that is obtained from a captured image does not give any insight into the actual performance. For example, capturing a defocused image with an extremely high resolution will yield a low effective resolution. We can determine and illustrate our sampling efficiency by generating a Modulation Transfer Function (MTF) curve. When the precision of our image sampling is determined using this method, we estimate its accuracy using objects with features at pre-determined positions. We will make a distinction between the in-plane dimensions (x, y) and out-of plane dimension z, as their performance significantly differs.

10-3-1 Precision and Effective Resolution

Determining the precision of the image sampling in an objective manner is a non trivial task. We defined the precision of our system as its effective resolution. This is when features can be distinguished with a confidence of approximately 68%.

In-Plane

We can estimate the effective in-plane resolution by looking at the labeled color that corresponds to each pixel, and investigate its transition. A complicating factor for our system is that the sampling is discrete. An intensity transition in the image that is perfectly sampled and sharp, generally lies on top of a sampled pixel. This means that this pixel samples half of each transition. In a transition from white to black, this pixel would be observed as gray, even though the edge is perfectly sharp and the image is accurately sampled. One such a transition is not sufficient to estimate the sampling efficiency, as the gray value might have been caused by defocus. We can obtain more measurement points by repeating this intensity transition measurement over a slanted edge, using a slanted edge target as depicted in Figure 10-5. If the target would be straight and not be slanted at an angle, we would repetitively get the same measurement values. Repeating the edge transition analysis over the slanted edge produces many values from which a fairly continuous edge transition line can be constructed. This edge-spread function is derived to a line-spread function, which is then binned and a discrete Fourier transform (DFT) is applied. If we normalize the result, we acquire the spatial frequency response (SFR) which gives us the MTF. This process is described and standardized by ISO norm 12233:2000.

From visual inspection of the shape of the MTF curve and derivation of some keypoints we can assess our precision. The slanted edge targets were captured by our system and processed, after which we used the RGB map for deriving the horizontal and vertical MTF curves. These curves are displayed in Figure 10-7. The performance metrics that are derived from these curves are depicted in Table 10-2. We can see that a distinction is made between vertical and horizontal performance, since we had vertical and horizontal slanted edges. Making this distinction is usually important in scanner cameras, where the data in one dimension is acquired by relative movement. However, this is also important for our setup, as both our cameras are rotated towards our plane of interest around the vertical axis. This means that our effective resolution is higher in the vertical direction (y) than in the horizontal direction (x). Because we have resampled our data with a fixed 2-D grid and a 22 µm resolution in both dimensions, we immediatelly see why we have made the distinction between sampled resolution and effective resolution. We can see that the sampling efficiency in vertical and horizontal directions is 48% and 34% respectively. The MTF-50% and MTF-10% metrics give the spatial frequency values at their respective modulation percentage. These give an indication of how our image deals with high and low frequency features. The sampling efficiencies can be multiplied by the resolutions to yield the effective resolutions [81] and are also stated in Table 10-2.

Our aim was for an in-plane resolution of 300 ppi (85 µm per pixel). Since our effective resolution is even lower than this value, we obtain a sampling effieiency at 300 ppi exceeding 100 %. This of course meets the requirements set by the FADGI guidelines for this resolution.

The fact that our sampling efficiency is not optimal with steps of 22 µm is apparent in the shape of the MTF curve. An intuitive graphical explanation of how to read an MTF curve is depicted in Figure 10-6. In this figure we can see how the MTF curve displays the modulation as a function of spatial frequency. In this example two binary patterns with different frequencies are photographed, yielding two points on the curve. The usage of a slanted edge is less intuitive, but will yield the entire curve. Note that a low modulation at a high frequency does not mean that image is not sharp, because the edges have been discretized as discussed before. The shape of our MTF curve looks very natural for an entirely unsharpened and raw image. The sampling efficiency can be artificially increased by sharpening the image, but such an operation will reveal itself as a bulge in the low frequency of the curve. The FADGI guidelines does not tolerate oversharpening and therefore set a threshold for the maximum MTF value of 1.0.

Out-Of-Plane

The in-plane (x and y dimensions) precision differs significantly from the out-of-plane (z) precision. Where the in-plane resolution is mostly defined by the size of the sensor and the object plane, the out-of-plane resolution is defined by the extracted phase value, which in turn is constructed by the modulation of the underlying fringes. The precision is therefore different and cannot be distinguished with an MTF curve.

We assess the depth precision by using the calibration plate in Figure 10-8. The calibration plate has been 3-D printed by using a layer deposition method and has cones of three defined

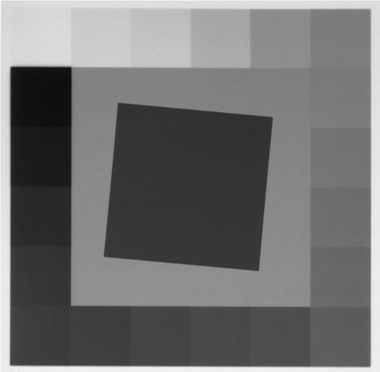


Figure 10-5: Slanted edge target 'QA62' used to measure the MTF response of the system.

Table 10-2: Spatial in-plane precision results.

	Horizontal (x)	Vertical (y)	Average
Spatial Resolution	22 µm	22 µm	22 µm
Sampling Efficiency	34 %	48 %	41 %
Effective Resolution	65 µm	46 µm	56 µm
MTF-10% (cy/mm)	9.42	13.3	11.36
MTF-50% (cy/mm)	4.43	4.44	4.44
Color Channel Misregistration (pixel)	0.02	0.06	0.04

heights. The different heights are approximately 1 mm, 2 mm and 3 mm and will be referred to as heights A, B and C respectively. The 3-D printing method stacks multiple sheets of plastic on top of each other. For our calibration plate, which means that the height of each separate cone height is similar. However, the height resolution is constrained by the layer thickness and will therefore not be precise. The calibration plate's use for assessing precision proved limited, because the layering was inconsistent and many layers that make up the cones had fallen off. This is why the classification between the three heights is made.

After scanning the plate, we sample the standard deviation or depth noise of each cone. We then make a distinction between the three different heights. The results are displayed in Table 10-3 and corresponding box-plots in Figure 10-9. These box-plots give an intuitive insight into the distribution of the measured data. Note that the deviation between measured heights as can be seen in Figure (a) is mostly due to the inaccurate deposition of layers, rather than a measurement error.

We can see that the precision is similar between the three height steps, which indicates that the precision is not depth dependent. Our results indicate that we should be able to resolve local features like canvas texture (deviates around 40 µm). In Table 10-3 we can see that we have height standard deviation of 9.2 µm, which means features of this size can be resolved with a certainty of approximately 68%.

Currently there exists no official guideline on the topographical digitization of paintings. Our depth resolution is sufficient in the sense that it is more precise than the in-plane resolution. Research still has to be done in investigating the depth resolution that is needed for a life-like topographical reproduction. Furthermore, cultural institutions have to get acquainted

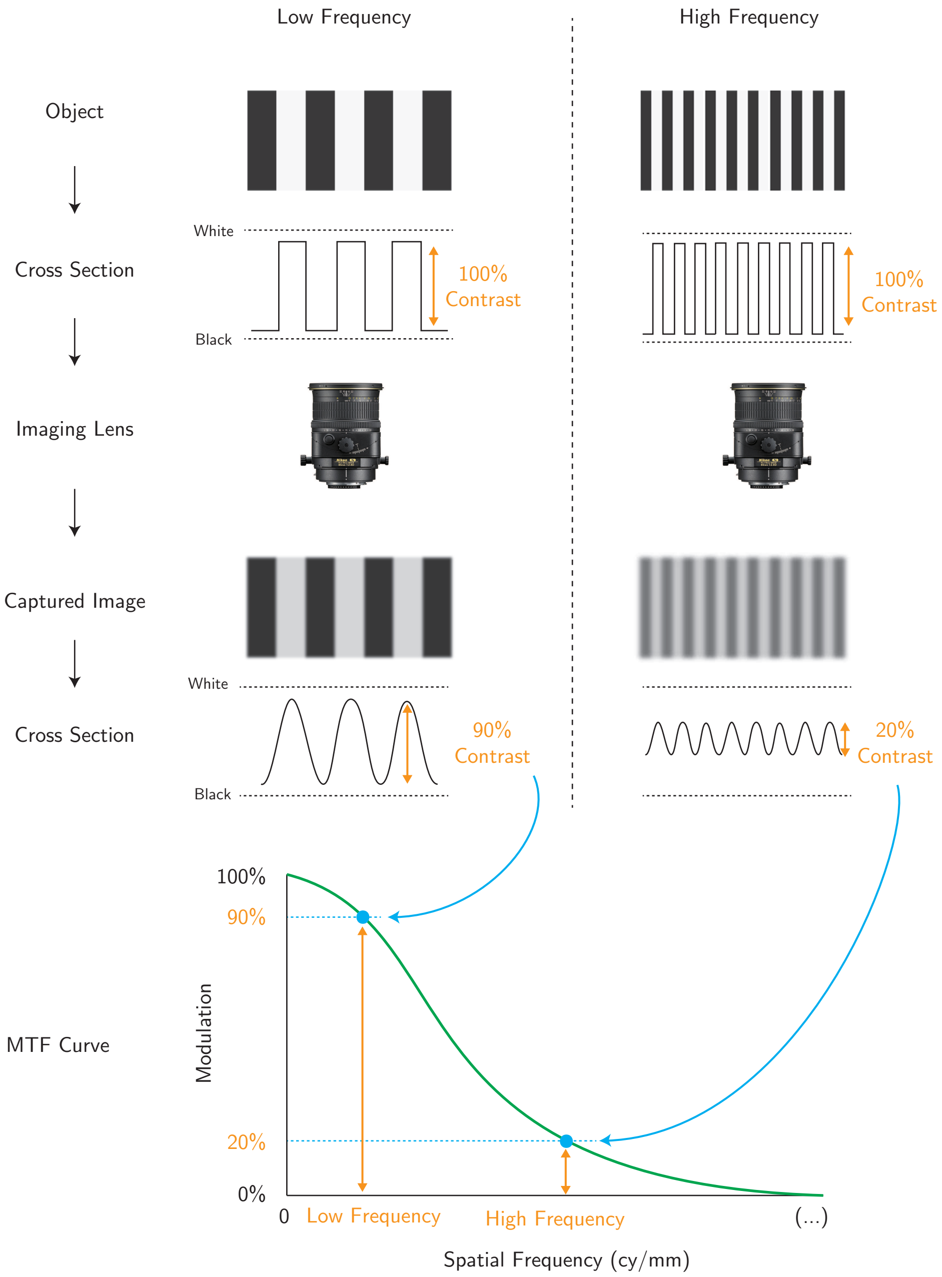
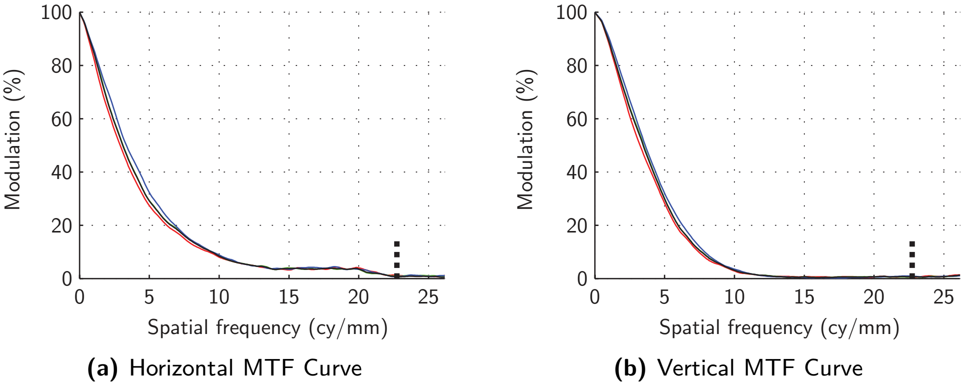


Figure 10-6: Intuitive explanation of the MTF curve featuring a low and a high frequency example.



(a) Horizontal MTF Curve

(b) Vertical MTF Curve

Figure 10-7: Modulation Transfer Function (MTF) results for the horizontal (a) and vertical (b) directions. The vertically dashed line indicates the half sampling frequency.

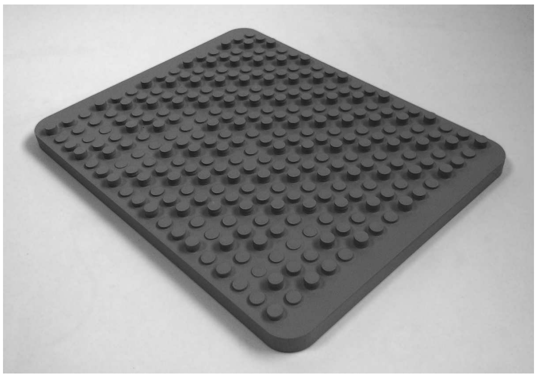


Figure 10-8: The calibration plate with cones with three different depths for out-of-plane spatial precision assessments.

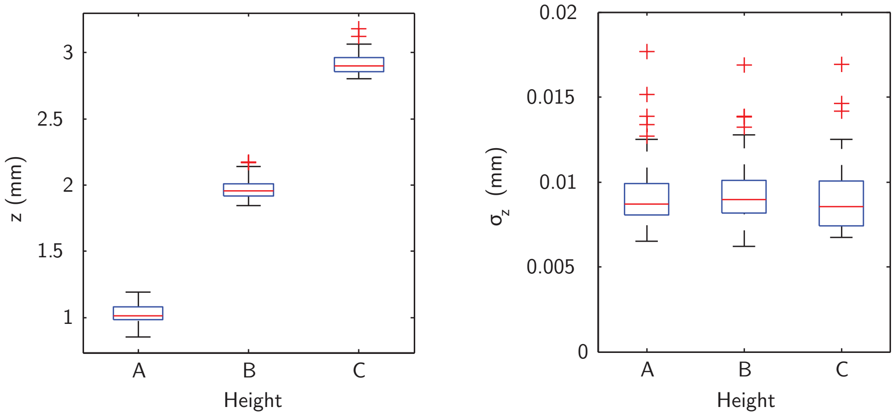
to depth maps as a new tool for research and what they can learn and distill from them. These can then relate such experiences to a preferred depth resolution, as was done with 2-D digitization in the form of the FADGI guidelines.

10-3-2 Accuracy

In order to perform accuracy tests, we use an object with known dimensions. For our purposes it is convenient to use an ideal planar surface through with we can measure the consistency of the z dimension. We can then use features with known positions such that these can be extracted from our data, and related to their respective locations. Corners of checkerboards

Table 10-3: Out-of-plane height (z) spatial precision measurements for three different height values using the depth performance target from Figure 10-8.

Height #	A	B	C	Average
Approximate real height	1 mm	2 mm	3 mm	
Height (z) std. deviation	9.4 µm	9.4 µm	8.9 µm	9.2 µm



(a) Measured Absolute Heights

(b) Local Height Deviation

Figure 10-9: Box-plots of the measured height of the cones in (a) of the calibration plate in Figure 10-8. In (b) the local (inter-cone) standard deviations are displayed that give an indication of the noise on pixel-level.

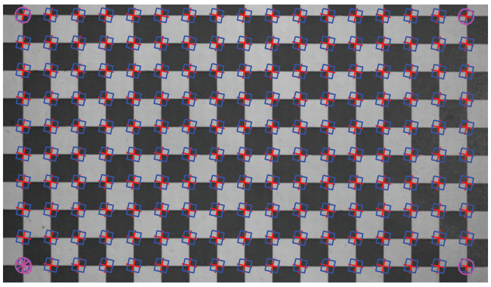


Figure 10-10: Captured RGB map of the checkerboard plane showing the localized corners.

are ideal for feature extraction, because they are easy to locate for algorithms and a highly accurate localization can be obtained. Therefore we use a printed checkerboard mounted on a flat surface. We then proceed to locate all the corners of the checkered patches as shown in Figure 10-10, yielding their measured x and y locations. Their z location is obtained from sampling the center of each point in the depth map. These obtained locations then have to be related to their real 3-D locations, from which we know that the checkers have a fixed size, are equally spaced apart and are mounted on a plane. The absolute Euclidean distance Δd is the accuracy metric as measured from our captured location to its corresponding absolute point on our fixed grid plane. The accuracy results are shown in a box-plot per dimension in Figure 10-11 and indicate the variance of the accuracy over the different measured locations. The data set is made up out of the Δd values per extracted corner. The mean accuracy values for each dimension are stated in Table 10-4. Similar to the precision results, observe that the accuracy in the horizontal direction (x) is worse compared to the vertical direction (y). The accuracy values are in the same order of magnitude as their respective precision results, except for the depth dimension (z). Its local precision is very high (9.2 µm) compared to its accuracy (38 µm). This is mainly because the precision in this dimension is dictated by the accurate modulation of the fringes. However, its accuracy it is still just as dependent on the orientation of our system and optical models as the other dimensions are (x and y).

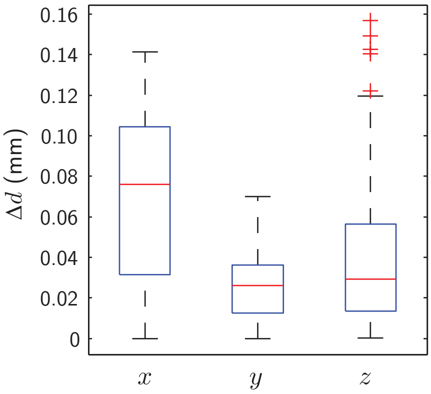


Figure 10-11: Box-plot showing depth accuracy Δd for the different dimensions in our checkered plane.

Table 10-4: Out-of-plane height (z) spatial precision measurements for three different height values using the depth performance target from Figure 10-8.

	x	y	z
Mean Accuracy Δd	68 µm	26 µm	38 µm

10-4 Speed and Size

The dimensions of the plane to be scanned without moving the system are dependent on the translation axis and the tripod it is mounted on. Currently, our (x) translation axis has a range of 110 cm, and the tripod's (y) range is 100 cm. The size can be further increased by moving the entire system along the plane. The capture of one single scene is 17 cm in x direction and 10 cm in the y direction, and sampled at 22 µm per pixel with an average sampling efficiency of 41%. The out-of-plane depth of field is around 10 mm, and the in-plane depth of field is unlimited due to the Scheimpflug lenses (given that the object stays parallel with the plane of focus).

A typical capture run of one scene consisting of 12 projections (yielding 24 images) takes 150 s. This is largely dependent on the transfer speed of the images from the camera to the computer, as our camera is rated to 6 frames per second at 40 megapixel, theoretically achieving a capture speed of within 3 seconds. The projector's refresh rate is not the bottleneck with a 60 Hz refresh rate.

Although one scene is captured in a fairly long 150 seconds, it has captured around 1 GB of raw image data, which is then converted to a usable 16-bit TIFF image format, yielding 6.4 GB. All the data was saved on solid state drives (SSD's) through a USB-3.0 interface.

The mathematical computations involved in fringe projection and stereo matching for images of this size and amount take up a lot of time. The peak RAM memory usage in one computational run is around 20 GB. The total computation time per scene was optimized down from 20 days per scene in the first implementation to currently around 600 to 700 seconds. For one scene, around 25 million 3-D points are triangulated (including RGB information). The speed per second is typically around 37,000 points per second on a single thread on a 3.4 GHz processor and a memory speed of 1.6 GHz.

In an overview of different methods by Salvi et al. [10], 6 structured light methods were compared (Monks et al., Posdamer et al., Guhring, Pribanic et al., Car and Hummel). In their results, they benchmarked the speed and constructed 3-D points per technique. These were respectively 307, 789, 1.993, 1.543 and 1.346 points per second. This was done on a 3.0 GHz processor. This indicates that our implementation of the hybrid stereo and fringe processing method is highly efficient.

Each capture yields an area of 136 cm. For large paintings up to 1 m, we need to take multiple captures with some partial overlap, resulting in around 100 images for such a size. Since one capture takes 150 seconds, 1 m will then take around 5 hours (accounting for overhead). This falls comfortably inside the time-span of a day or night of scanning in a museum.

10-5 Artifacts and Exceptions

Although the test and calibration charts we have previously used possess some of the characteristics that paintings have, they are in fact not paintings. We therefore perform tests in different conditions to see what effect it has on the final image.

Varnished paintings can be highly reflective. These reflections can be entirely suppressed by the cross-polarization setup, which results in images shown in Figure 10-12. Images that do have specular reflections can still be processed, but overexposure will lead to clipping. The fringe processing will be inaccurate and we will observe artifacts. Color information is also lost at these locations.

Different environmental lighting should not be a problem, as this can be seen as an offset component that is present in each fringe, which does not influence the computation. Problems can arise when this offset component is different within fringes that belong to the same capture. This can be occur due to turning off lights, clouds passing by or shadows of bystanders. The latter is still not a problem when the difference between our illumination (the projector) and the environmental lighting is large. When the observed projector's brightness is impaired because of the cross-polarization filters, it is generally a good idea to avoid environmental lighting.

The low light intensity of our projector and the high aperture value on the camera's optics meant that we need a long exposure time (seconds). This meant that the system is prone to perturbations, since it is not very sturdy due to its portability, although in practice this proved to be no problem on a solid floor.

Through color-patched charts, we had already verified that we can capture color data. To make sure these different colors are correctly processed and do not induce a bias in our depth information, a special full color 3-D print was used as displayed in Figure 10-13. This print contains six colors printed on a white surface. In the depth map we can clearly see that the depth profile of each stroke is triangular, but the angle changes with the length of the stroke. We can see that each color is properly processed. We then investigated what happened when the angle of the 3-D printed strokes got very steep. We can see in Figure 10-14 that the system generally has difficulties producing enough depth data when this angle is between 90° (perpendicular to the surface) and 45°. The black regions indicate missing data. This fact comes to no surprise as the angle between the two cameras is around 40° and impairs

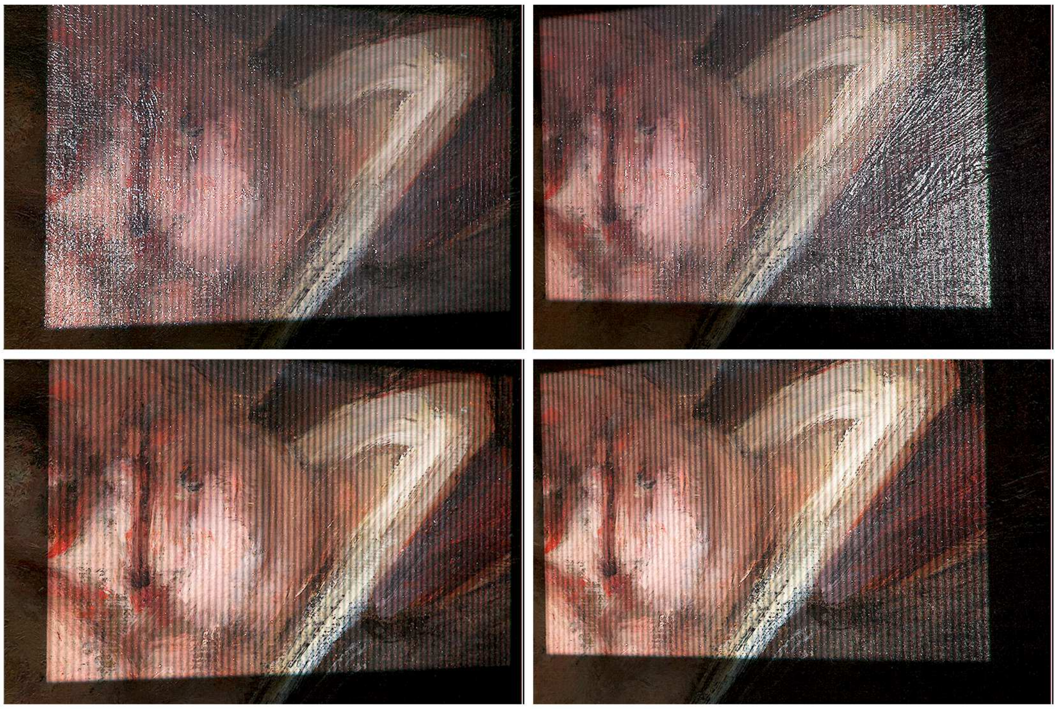


Figure 10-12: The columns depict images taken of a fringe from the left and right cameras respectively. In the bottom row, cross polarization is used, whereas in the top row only single polarization filters are used. Specular reflections that arose from the thick layer of varnished proved to degrade the accuracy of the system. Cross-polarization method was therefore used to suppress all reflections.

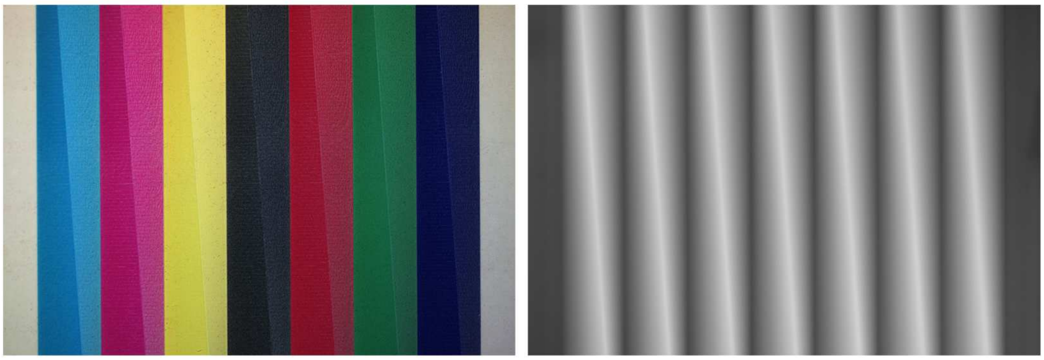


Figure 10-13: Capture showing the system's indifference to color. This 3-D print was made using identical 3-D structures with various colors. We can see that the depth map (right) illustrates the same behavior for each color.

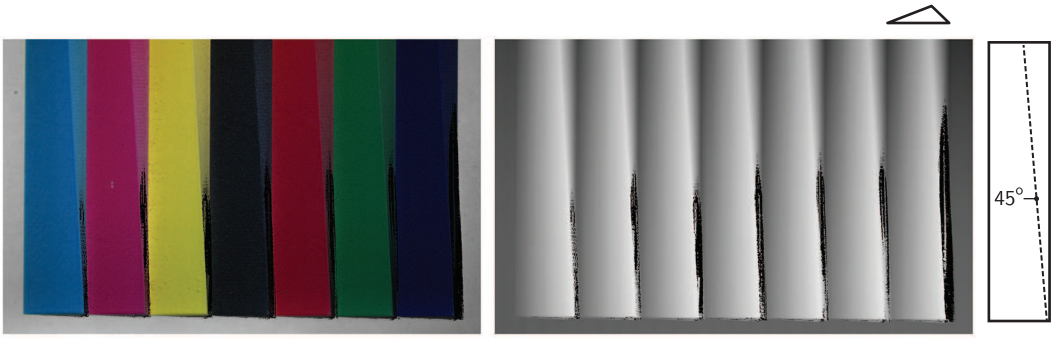


Figure 10-14: Edge the object shown in Figure 10-13. In this object, the angle of the 3-D structure with respect to the plane it is printed on is gradually increased. In the depth map (right) we can see that our system rejects much data from slopes steeper than $45°$ because of too low confidence values.

observing corresponding points at such steep angles. This condition only holds for vertical edges.

10-6 Overview

The color reproduction properties and in-plane spatial accuracy of the system is sufficient as defined by the FADGI guidelines for the digitization of cultural heritage. The obtained depth information is very precise, which enables us to obtain relevant features as small as the texture of the canvas. One scene is captured in roughly 2.5 minutes. The efficient data processing converts the large amount of images to a point cloud of between 20 and 30 million points, each labeled with color information. These metrics are all summarized in Table 10-5. From the computed point-cloud, a depth map and RGB map are generated that can immediately be used as input to the full-color 3-D printer. An example is illustrated in Figure 10-15 where we can see an RGB map and a depth map. A large scale image of this same painting is depicted in Figure 10-19. In this document, we have not focused on image enhancement, noise correction or other types of filtering. As an example of what is possible in that domain, we

Table 10-5: Overview of the system's key performance indicators.

	x	y	z
Size	17 cm	10 cm	1 cm
Sampled Resolution	22 µm	22 µm	N/A
Effective Resolution	65 µm	46 µm	9.2 µm
Accuracy	68 µm	26 µm	38 µm
Color Accuracy	1.9 ΔE-2000		
Color Noise	0.39 %		
Capture Speed	150 s		
Processing Speed	650 s		

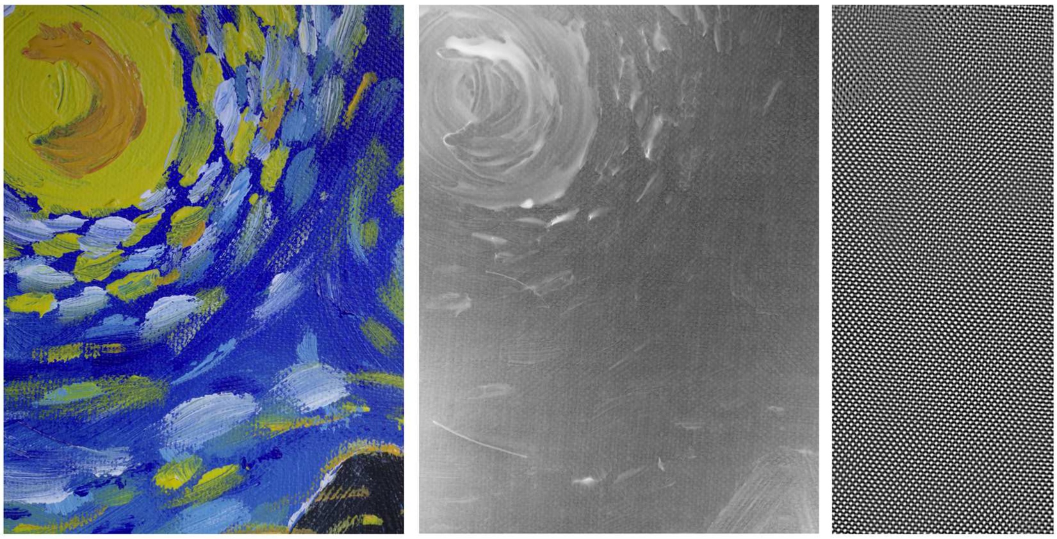


Figure 10-15: Detail of the painting *Ceci n'est pas une reproduction* by Tim Zaman (acrylic on canvas, 2013). On the left and middle we see a RGB map and depth map, respectively. On the right, the texture of the canvas was extracted using Fourier filtering.

have extracted the fine harmonics of the canvas structure in the Fourier domain as depicted in Figure 10-15. This way we can subtract the canvas structure from the depth map, as its structure is often irrelevant to the painting as the artist envisioned it. The 3-D printing process can then be done on a real canvas with its own structure, only printing the structure of the paint.

10-7 Practical Test Case

A painting of a self-portrait by Rembrandt was scanned using our system on the 18th of February. The painting is owned by the Mauritshuis, and was scanned at the Gemeentemuseum in The Hague. It was painted with oil on canvas in the year 1669 and has a height of 65 cm and a width of 60 cm. Photos of the setup are depicted in Figure 10-16. The capture of the entire painting consisted of 40 captures, and took 3.5 hours. Setup and calibration

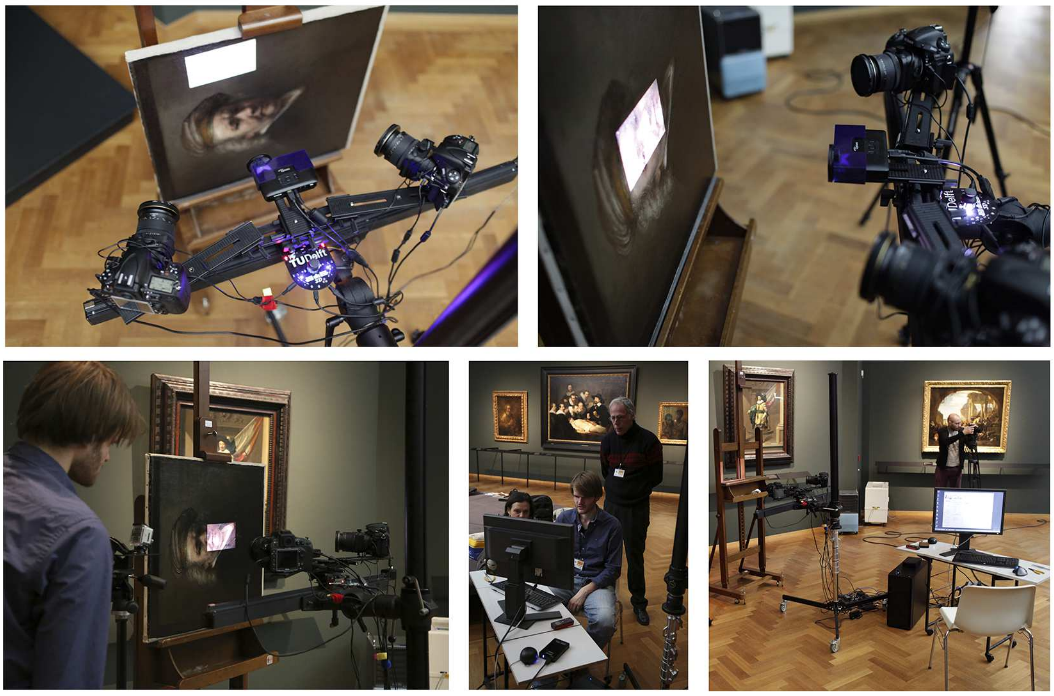


Figure 10-16: Photos taken of the setup during the scanning of the painting by Rembrandt.

took around 2 hours, and capturing test images another hour. The painting has a dark appearance and was heavily varnished. This meant that cross-polarization was necessary. The brightness of the small projector proved low compared to the environmental lighting, which were therefore turned off. Two details from different captures are depicted in Figure 10-17 and Figure 10-18. In these figures, the constrast has been enhanced of this document, no other post-processing has been conducted. In both images, the maximum depth deviation is around 0.5 mm.

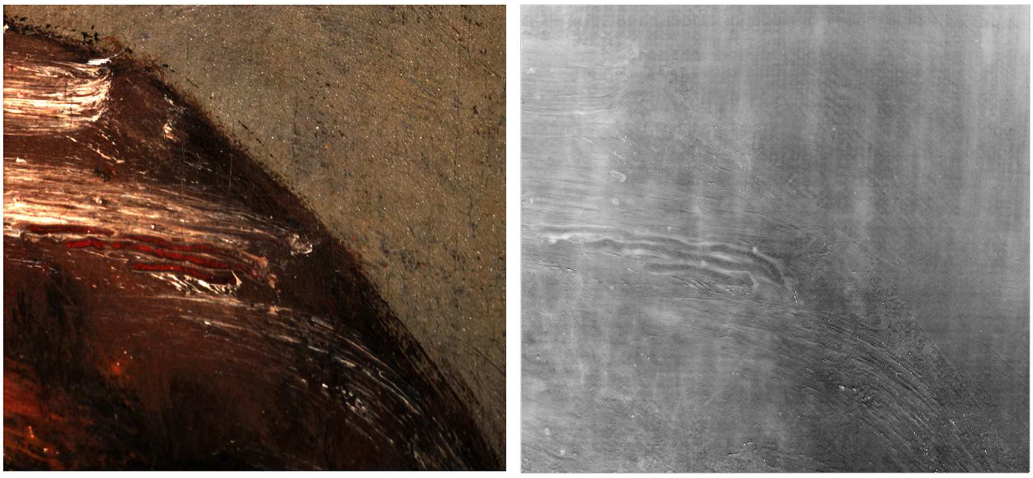


Figure 10-17: Detail of three lines Rembrandt scratched into the wet paint of his self-portrait. In the depth map (right) the three scratches can easily be distinguished, including the sides of the scratches where the paint accumulated.

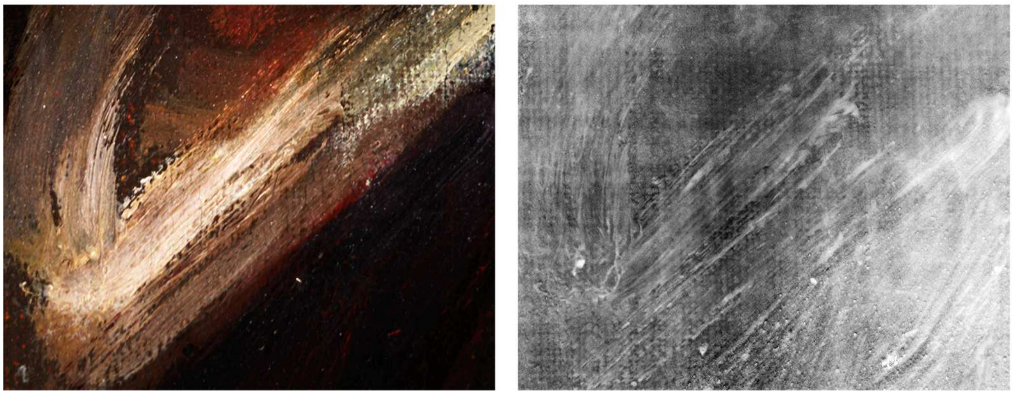


Figure 10-18: Detail of the collar of the self-portrait by Rembrandt. In the depth map, we can see that the single dash of white paint is fairly thick and covers up the canvas behind it. We can also see that the black paint layer is thicker and is granular. Note that the contrast has been greatly increased.

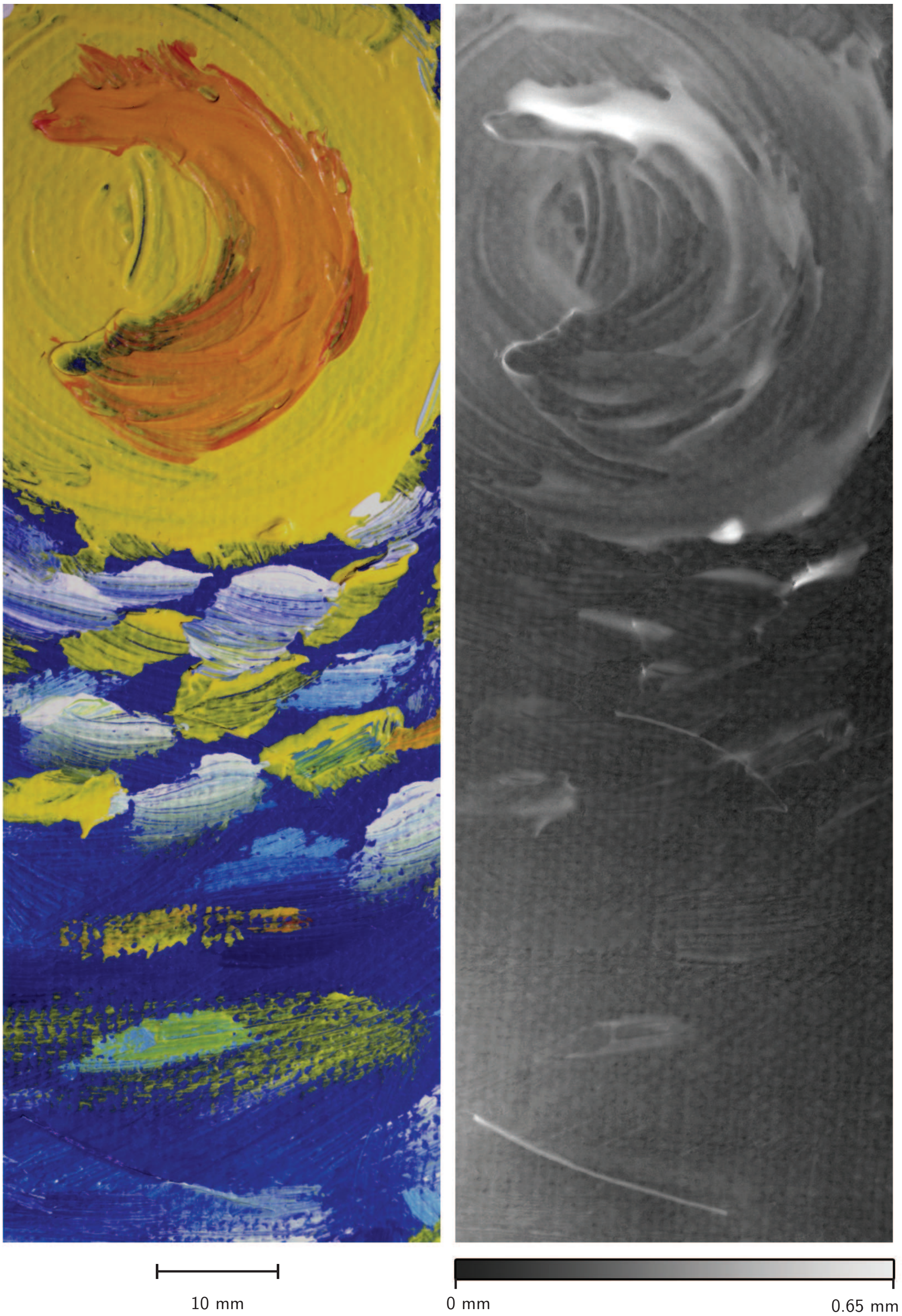


Figure 10-19: Large scale detail of the RGB map and depth map of the image in Figure 10-15.

Chapter 11

Conclusion

Paintings are versatile near-planar objects that have material characteristics that vary widely. The fact that paint has a material presence is often overlooked, mostly because of the fact that we encounter many of these artworks through two dimensional reproductions. The capture of paintings in the third dimension is not only interesting for study, restoration and conservation, but it also facilitates making three dimensional reproductions through novel 3-D printing methods. In order to print in full color as well, the color of the painting has to be captured. An overview was made of the feasible 3-D imaging methods that can accommodate the accurate capture of the painting's surface. The lack of an ideal solution for objects as versatile as paintings, lead to the construction of a hybrid system. Fringe projection and stereo imaging are combined to yield an accurate, fast and reliable method. The usage of high-end cameras, special lenses and filters accommodate the capture of unparalleled amounts of 3-D data per capture, while simultaneously capturing color information. A printed circuit board was made to control the hardware and automating the translation stage, which allows for capturing planes measuring up to 1 meter.

We have demonstrated that fringe projection is a very effective method of encoding images independent of the projector's resolution. The projection facilitates that each pixel observed in one camera of a stereo setup can be easily matched to a pixel in the other camera. Through our combined camera and projector calibration, we can use these pixel locations to triangulate an absolute 3-D point in space. A prerequisite of using fringe projection as image encoder is that both observed fringe images are correlated to each other. This was done through sparse stereo matching. The consequent triangulation of around 30 million data points per capture was done through a special look-up table construction and by using pre-computed optical ray maps for each camera sensor. The processing takes around 650 s, while the capture of one scene takes 150 s. The reproductive performance for both color and spatial properties were satisfactory related to digitization guidelines. We obtain an in-plane precision of around 50 µm when capturing an area of 170 cm^2. An effective depth resolution of 9.2 µm then allows for resolving the smallest visible features in a painting.

The first results from the capture of a painting by Rembrandt using our design indicate that the system works well in practice. The depth map and color image resulting from this will be used for the construction of a reproduction in full color and full dimension. Our ever increasing ability of scanning and reproducing paintings will eventually result in the ultimate reproduction, which redefines the answer to the first question in this document:

What is a painting?

11-1 Improvements and Future Work

In this document we have focused primarily on the selection, refinement and fusion of the best 3-D imaging technique for paintings. The fact that we have designed such a device that is ideally suited to do so, does not exclude the viability of other methods. Requirements around the 3-D scanning of paintings still remain to be set. Studies have to be carried out into which performance is needed for the purpose of study, conservation and restoration, but also from reproduction. Such solid requirements are important in the design of an ideal system.

The design is mounted on a horizontal translation stage. In order to capture large paintings efficiently and position itself more accurately, the vertical axis should be automated as well. The construction should be more rigid with a heavier support and translation stage. The cameras and projector should be fixed in a rigid mount. The projector can then be upgraded to emit more light intensity.

Surfaces that reflect an equal light intensity in all directions do not appear equally bright when viewed from a varying angle (Lambert's cosine law). We have not taken this fact into account in this design, although it should be a straightforward implementation. In practice this means that slopes often appear brighter or darker than their surroundings, although they might have the same color. We wish to eliminate this in our color images, because this phenomenon should reveal itself automatically through the 3-D print.

Bibliography

[1] M. Levoy, "The digital michelangelo project," in *3-D Digital Imaging and Modeling, 1999. Proceedings. Second International Conference*, pp. 2–11, IEEE, 1999.

[2] W. Taft and J. Mayer, *The Science of Paintings.* Springer, 2001.

[3] S. Reichelt, R. Häussler, G. Fütterer, and N. Leister, "Depth cues in human visual perception and their realization in 3d displays," in *Proc. SPIE*, vol. 7690, p. 76900B, 2010.

[4] J. Caarls, *Pose estimation for mobile devices and Augmented Reality.* PhD thesis, Delft University of Technology, 2009.

[5] J. Cutting and P. Vishton, "Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth," *Perception of space and motion*, vol. 5, pp. 69–117, 1995.

[6] E. Adelson and J. Wang, "Single lens stereo with a plenoptic camera," *IEEE transactions on pattern analysis and machine intelligence*, vol. 14, no. 2, pp. 99–106, 1992.

[7] X. Chen, J. Xi, Y. Jin, and J. Sun, "Accurate calibration for a camera–projector measurement system based on structured light projection," *Optics and Lasers in Engineering*, vol. 47, no. 3, pp. 310–319, 2009.

[8] S. Gorthi and P. Rastogi, "Fringe projection techniques: Whither we are?," *Optics and Lasers in Engineering*, vol. 22, no. 2, p. 133, 2009.

[9] C. Quan, W. Chen, and C. Tay, "Phase-retrieval techniques in fringe-projection profilometry," *Optics and Lasers in Engineering*, vol. 48, no. 2, pp. 235–243, 2010.

[10] J. Salvi, S. Fernandez, T. Pribanic, and X. Llado, "A state of the art in structured light patterns for surface profilometry," *Pattern recognition*, vol. 43, no. 8, pp. 2666–2680, 2010.

[11] J. Wassink, "Vijf TU-onderzoekers ontvangen Vidi-beurs NWO," *TU Delta*, vol. 42-34.

[12] T. Zaman, “Moving Image Quality Analysis to the Cloud,” in *Proceedings of IS&T Archiving 2012*, (Copenhagen, Denmark), pp. 183–185, July 2012.

[13] L. Capell, “Digitization as a preservation method for damaged acetate negatives: A case study,” *American Archivist*, vol. 73, no. 1, pp. 235–249, 2010.

[14] R. Newman, “Some applications of infrared spectroscopy in the examination of painting materials,” *Journal of the American Institute for Conservation*, pp. 42–62, 1979.

[15] A. Adam, P. Planken, S. Meloni, and J. Dik, “TeraHertz imaging of hidden paint layers on canvas,” in *Infrared, Millimeter, and Terahertz Waves, 2009. IRMMW-THz 2009. 34th International Conference*, pp. 1–2, IEEE, 2009.

[16] J. Dik, K. Janssens, G. Van Der Snickt, L. Van Der Loeff, K. Rickers, and M. Cotte, “Visualization of a lost painting by Vincent van Gogh using synchrotron radiation based X-ray fluorescence elemental mapping,” *Analytical chemistry*, vol. 80, no. 16, pp. 6436–6442, 2008.

[17] S. Lang, “Method for reproducing paintings and the like,” Nov. 20 1990. US Patent 4,971,743.

[18] S. Lang and H. Kalef, “Method and means of publishing images having coloration and three-dimensional texture,” Jan. 26 1993. US Patent 5,182,063.

[19] S. Hussmann, T. Ringbeck, and B. Hagebeuker, “A performance review of 3d tof vision systems in comparison to stereo vision systems,” *Stereo Vision*, pp. 103–120, 2008.

[20] S. May, B. Werner, H. Surmann, and K. Pervolz, “3d time-of-flight cameras for mobile robotics,” in *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference*, pp. 790–795, IEEE, 2006.

[21] P. Treleaven and J. Wells, “3d body scanning and healthcare applications,” *Computer*, vol. 40, no. 7, pp. 28–34, 2007.

[22] T. Nguyen, T. Qui, K. Xu, A. Cheok, S. Teo, Z. Zhou, A. Mallawaarachchi, S. Lee, W. Liu, H. Teo, *et al.*, “Real-time 3d human capture system for mixed-reality art and entertainment,” *Visualization and Computer Graphics, IEEE Transactions*, vol. 11, no. 6, pp. 706–721, 2005.

[23] T. Moeslund, A. Hilton, and V. Krüger, “A survey of advances in vision-based human motion capture and analysis,” *Computer vision and image understanding*, vol. 104, no. 2, pp. 90–126, 2006.

[24] C. Quan, C. Tay, X. He, X. Kang, and H. Shang, “Microscopic surface contouring by fringe projection method,” *Optics & Laser Technology*, vol. 34, no. 7, pp. 547–552, 2002.

[25] A. Yao, “Applications of 3d scanning and reverse engineering techniques for quality control of quick response products,” *The International Journal of Advanced Manufacturing Technology*, vol. 26, no. 11, pp. 1284–1288, 2005.

[26] W. Carter, R. Shrestha, and K. Slatton, “Geodetic laser scanning,” *Physics Today*, vol. 60, p. 41, 2007.

[27] F. Bernardini, H. Rushmeier, I. Martin, J. Mittleman, and G. Taubin, "Building a digital model of Michelangelo's Florentine Pieta," *Computer Graphics and Applications, IEEE*, vol. 22, no. 1, pp. 59–67, 2002.

[28] G. Guidi, M. Pieraccini, S. Ciofi, V. Damato, J. Beraldin, and C. Atzeni, "Tridimensional digitizing of Donatello's Maddalena," 2001.

[29] G. Sansoni and F. Docchio, "3-d optical measurements in the field of cultural heritage: the case of the Vittoria Alata of Brescia," *Instrumentation and Measurement, IEEE Transactions*, vol. 54, no. 1, pp. 359–368, 2005.

[30] C. Christofides, B. Castellon, Object, K. Polikreti, and C. de Deyne, "Fine art painting characterization by spectroscopic ellipsometry: preliminary measurements on varnish layers," *Thin Solid Films*, vol. 455-456, pp. 207 – 212, 2004. The 3rd International Conference on Spectroscopic Ellipsometry.

[31] E. Berry, "Diffuse reflection of light from a matt surface," *JOSA*, vol. 7, no. 8, pp. 627–633, 1923.

[32] S. Nayar, X. Fang, and T. Boult, "Separation of reflection components using color and polarization," *International Journal of Computer Vision*, vol. 21, no. 3, pp. 163–186, 1997.

[33] "Camera polarizing filters." http://www.cambridgeincolour.com/tutorials/polarizing-filters.htm. Accessed: 14/09/2012.

[34] S. Nayar and M. Gupta, "Diffuse structured light," in *Computational Photography (ICCP), 2012 IEEE International Conference*, pp. 1–11, IEEE, 2012.

[35] F. de Jong, *Range Imaging and Visual Servoing for Industrial Applications.* PhD thesis, Delft University of Technology, 2008.

[36] S. Staniforth, "Retouching and colour matching: the restorer and metamerism," *Studies in conservation*, pp. 101–111, 1985.

[37] W. Benjamin, *The Work of Art in the Age of Mechanical Reproduction.* Penguin Books Limited, 2008.

[38] B. Rogers, M. Graham, *et al.*, "Motion parallax as an independent cue for depth perception," *Perception*, vol. 8, no. 2, pp. 125–134, 1979.

[39] D. Mitchell, "Retinal disparity and diplopia," *Vision research*, vol. 6, no. 7, pp. 441–451, 1966.

[40] H. Hecht, U. B. Bielefeld, M. K. Kaiser, and M. S. Banks, "Gravitational acceleration as a cue for absolute size and distance," *Psychophysics*, pp. 1066–1075, 1996.

[41] T. Stroffregen and J. Pittenger, "Human echolocation as a basic form of perception and action," *Ecological psychology*, vol. 7, no. 3, pp. 181–216, 1995.

[42] D. Knill *et al.*, "Mixture models and the probabilistic structure of depth cues," *Vision research*, vol. 43, no. 7, pp. 831–854, 2003.

[43] R. Jacobs, “Optimal integration of texture and motion cues to depth,” *Vision research*, vol. 39, no. 21, pp. 3621–3629, 1999.

[44] H. Bülthoff and H. Mallot, “Integration of depth modules: stereo and shading,” *JOSA A*, vol. 5, no. 10, pp. 1749–1758, 1988.

[45] M. Johnson and E. Adelson, “Retrographic sensing for the measurement of surface texture and shape,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference*, pp. 1070–1077, IEEE, 2009.

[46] J. Adams, K. Parulski, and K. Spaulding, “Color processing in digital cameras,” *Micro, IEEE*, vol. 18, no. 6, pp. 20–30, 1998.

[47] T. Scheimpflug, “Improved method and apparatus for the systematic alteration or distortion of plane pictures and images by means of lenses and mirrors for photography and for other purposes,” *GB patent*, vol. 1196, 1904.

[48] R. Lange and P. Seitz, “Solid-state time-of-flight range camera,” *Quantum Electronics, IEEE Journal of*, vol. 37, no. 3, pp. 390–397, 2001.

[49] S. Gudmundsson, H. Aanaes, and R. Larsen, “Fusion of stereo vision and time-of-flight imaging for improved 3d estimation,” *International Journal of Intelligent Systems Technologies and Applications*, vol. 5, no. 3, pp. 425–433, 2008.

[50] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, “Light field photography with a hand-held plenoptic camera,” *Computer Science Technical Report CSTR*, vol. 2, 2005.

[51] T. Georgiev, G. Chunev, and A. Lumsdaine, “Superresolution with the focused plenoptic camera,” in *IS&T/SPIE Electronic Imaging*, pp. 78730X–78730X, International Society for Optics and Photonics, 2011.

[52] M. Kimura, M. Mochimaru, and T. Kanade, “Projector calibration using arbitrary planes and calibrated camera,” in *Computer Vision and Pattern Recognition, 2007. CVPR’07. IEEE Conference*, pp. 1–2, IEEE, 2007.

[53] M. Levoy, B. Chen, V. Vaish, M. Horowitz, I. McDowall, and M. Bolas, “Synthetic aperture confocal imaging,” in *ACM Transactions on Graphics (TOG)*, vol. 23, pp. 825–834, ACM, 2004.

[54] A. Pentland, “A new sense for depth of field,” *Pattern Analysis and Machine Intelligence, IEEE Transactions*, no. 4, pp. 523–531, 1987.

[55] J. Elder and S. Zucker, “Local scale control for edge detection and blur estimation,” *Pattern Analysis and Machine Intelligence, IEEE Transactions*, vol. 20, no. 7, pp. 699–716, 1998.

[56] H. Hu and G. de Haan, “Low cost robust blur estimator,” in *Image Processing, 2006 IEEE International Conference*, pp. 617–620, IEEE, 2006.

[57] B. Horn, “Obtaining shape from shading information,” in *Shape from shading*, pp. 123–171, MIT Press, 1989.

[58] R. Woodham, "Photometric method for determining surface orientation from multiple images," *Optical engineering*, vol. 19, no. 1, pp. 191139–191139, 1980.

[59] M. Johnson, F. Cole, A. Raj, and E. Adelson, "Microgeometry capture using an elastomeric sensor," *ACM Transactions on Graphics (TOG)*, vol. 30, no. 4, p. 46, 2011.

[60] J. Salvi, J. Pages, and J. Batlle, "Pattern codification strategies in structured light systems," *Pattern Recognition*, vol. 37, no. 4, pp. 827–849, 2004.

[61] J. Batlle, E. Mouaddib, and J. Salvi, "Recent progress in coded structured light as a technique to solve the correspondence problem: a survey," *Pattern recognition*, vol. 31, no. 7, pp. 963–982, 1998.

[62] M. Gupta, A. Agrawal, A. Veeraraghavan, and S. Narasimhan, "Structured light 3d scanning in the presence of global illumination," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference*, pp. 713–720, IEEE, 2011.

[63] M. Gupta, A. Agrawal, A. Veeraraghavan, and S. Narasimhan, "A practical approach to 3d scanning in the presence of interreflections, subsurface scattering and defocus," *International Journal of Computer Vision*, pp. 1–23, 2012.

[64] S. Gorthi and P. Rastogi, "Fringe projection techniques: whither we are?," *Optics and lasers in engineering*, vol. 48, no. 2, pp. 133–140, 2010.

[65] J. Pan, P. Huang, and F. Chiang, "Color-coded binary fringe projection technique for 3-d shape measurement," *Optical Engineering*, vol. 44, no. 2, pp. 023606–023606, 2005.

[66] M. Halioua, R. Krishnamurthy, H. Liu, and F. Chiang, "Projection moiré with moving gratings for automated 3-d topography," *Applied Optics*, vol. 22, no. 6, pp. 850–855, 1983.

[67] M. Takeda, H. Ina, and S. Kobayashi, "Fourier-transform method of fringe-pattern analysis for computer-based topography and interferometry," *JosA*, vol. 72, no. 1, pp. 156–160, 1982.

[68] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference*, vol. 1, pp. I–195, IEEE, 2003.

[69] H. Nishihara, "Practical real-time imaging stereo matcher," *Readings in computer vision*, pp. 63–72, 1987.

[70] L. Zhang, B. Curless, and S. Seitz, "Spacetime stereo: Shape recovery for dynamic scenes," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference*, vol. 2, pp. II–367, IEEE, 2003.

[71] C. Rocchini, P. Cignoni, C. Montani, P. Pingi, and R. Scopigno, "A low cost 3d scanner based on structured light," in *Computer Graphics Forum*, vol. 20, pp. 299–308, Wiley Online Library, 2001.

[72] D. Akca, A. Gruen, B. Breuckmann, C. Lahanier, D. Akca, D. Akca, A. Grün, and A. Grün, *High definition 3D-scanning of arts objects and paintings.* Institute of Geodesy and Photogrammetry, ETH Zurich, 2007.

[73] R. Fontana, M. Gambino, E. Pampaloni, L. Pezzati, and C. Seccaroni, "Panel painting surface investigation by conoscopic holography," in *8th International Conference on Non Destructive Investigations and Microanalysis for the Diagnostics and the Conservation of the Cultural and Environmental Heritage, Lecce*, pp. 15–19, 2005.

[74] G. Sansoni, M. Trebeschi, and F. Docchio, "State-of-the-art and applications of 3d imaging sensors in industry, cultural heritage, medicine, and criminal investigation," *Sensors*, vol. 9, no. 1, pp. 568–601, 2009.

[75] G. Pavlidis, A. Koutsoudis, F. Arnaoutoglou, V. Tsioukas, and C. Chamzas, "Methods for 3d digitization of cultural heritage," *Journal of Cultural Heritage*, vol. 8, no. 1, pp. 93–98, 2007.

[76] M. Pieraccini, G. Guidi, and C. Atzeni, "3d digitizing of cultural heritage," *Journal of Cultural Heritage*, vol. 2, no. 1, pp. 63–70, 2001.

[77] W. Boehler and A. Marbs, "3d scanning and photogrammetry for heritage recording: a comparison," in *Proceedings of the 12th International Conference on Geoinformatics*, pp. 291–298, 2004.

[78] W. Boehler, M. Bordas Vicent, and A. Marbs, "Investigating laser scanner accuracy," *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 34, no. Part 5, pp. 696–701, 2003.

[79] C. Curcio, K. Sloan, R. Kalina, and A. Hendrickson, "Human photoreceptor topography," *The Journal of comparative neurology*, vol. 292, no. 4, pp. 497–523, 1990.

[80] T. Poggio, M. Fahle, and S. Edelman, "Fast perceptual learning in visual hyperacuity," *Science*, vol. 256, no. 5059, pp. 1018–1021, 1992.

[81] P. Burns and D. Williams, "Sampling efficieny in digital camera performance standards," in *Proc. SPIE*, vol. 6808, 2008.

[82] US Federal Agencies Digitization Initiative Still Image Working Group, *Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files*, 2010.

[83] "Lighting for musea and expositions," tech. rep., NSVV workgroup Museumlighting, 2008.

[84] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International journal of computer vision*, vol. 47, no. 1, pp. 7–42, 2002.

[85] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[86] T. Kanade and M. Okutomi, "A stereo matching algorithm with an adaptive window: Theory and experiment," *Pattern Analysis and Machine Intelligence, IEEE Transactions*, vol. 16, no. 9, pp. 920–932, 1994.

[87] G. Sharma, *Digital Color Imaging Handbook.* Boca Raton, FL, USA: CRC Press, Inc., 2002.

[88] R. Tsai, "A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses," *Robotics and Automation, IEEE Journal of*, vol. 3, no. 4, pp. 323–344, 1987.

[89] Z. Zhang, "A flexible new technique for camera calibration," *Pattern Analysis and Machine Intelligence, IEEE Transactions*, vol. 22, no. 11, pp. 1330–1334, 2000.

[90] Y. Surrel, "Design of algorithms for phase measurements by the use of phase stepping," *Applied optics*, vol. 35, no. 1, pp. 51–60, 1996.

[91] P. Huang, Q. Hu, and F. Chiang, "Double three-step phase-shifting algorithm," *Applied optics*, vol. 41, no. 22, pp. 4503–4509, 2002.

[92] Y. Liu, "Accuracy improvement of 3d measurement using digital fringe projection," Master's thesis, University of Wollongong, 2011.

[93] J. Wyant, "Interferometric optical metrology: basic principles and new systems," *Laser Focus*, vol. 18, no. 5, pp. 65–71, 1982.

[94] P. Hariharan, B. Oreb, and T. Eiju, "Digital phase-shifting interferometry," *Appl Opt*, vol. 26, pp. 2504–2506, 1987.

[95] P. de Groot, "Long-wavelength laser diode interferometer for surface flatness measurement," in *Optics for Productivity in Manufacturing*, pp. 136–140, International Society for Optics and Photonics, 1994.

[96] S. Lei and S. Zhang, "Digital sinusoidal fringe pattern generation: defocusing binary patterns vs focusing sinusoidal patterns," *Optics and Lasers in Engineering*, vol. 48, no. 5, pp. 561–569, 2010.

[97] M. Tsai and C. Hung, "Development of a high-precision surface metrology system using structured light projection," *Measurement*, vol. 38, no. 3, pp. 236–247, 2005.

[98] G. Falcao, N. Hurtos, and J. Massich, "Plane-based calibration of a projector-camera system," *VIBOT Master*, 2008.

[99] A. Griesser and L. Van Gool, "Automatic interactive calibration of multi-projector-camera systems," in *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference*, pp. 8–8, IEEE, 2006.

[100] Z. Li, Y. Shi, C. Wang, and Y. Wang, "Accurate calibration method for a structured light system," *Optical Engineering*, vol. 47, no. 5, pp. 053604–053604, 2008.

[101] M. Ashdown and Y. Sato, "Steerable projector calibration," in *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference*, pp. 98–98, IEEE, 2005.

[102] J. Dijk, S. Verbeek, and Voor, "In search of an objective measure for the perceptual quality of printed images," 2004.

[103] S. Zhang and S. Yau, "Generic nonsinusoidal phase error correction for three-dimensional shape measurement using a digital video projector," *Applied optics*, vol. 46, no. 1, pp. 36–43, 2007.

[104] T. Coleman and Y. Li, "An interior trust region approach for nonlinear minimization subject to bounds," *SIAM journal on optimization*, vol. 6, no. 2, pp. 418–445, 1996.

[105] M. Herráez, D. Burton, M. Lalor, and M. Gdeisat, "Fast two-dimensional phase-unwrapping algorithm based on sorting by reliability following a noncontinuous path," *Applied Optics*, vol. 41, no. 35, pp. 7437–7444, 2002.

[106] M. Costantini, "A novel phase unwrapping method based on network programming," *Geoscience and Remote Sensing, IEEE Transactions*, vol. 36, no. 3, pp. 813–821, 1998.

[107] C. Quan, C. Tay, and L. Chen, "A study on carrier-removal techniques in fringe projection profilometry," *Optics & Laser Technology*, vol. 39, no. 6, pp. 1155–1161, 2007.

[108] S. Park, M. Park, and M. Kang, "Super-resolution image reconstruction: a technical overview," *Signal Processing Magazine, IEEE*, vol. 20, no. 3, pp. 21–36, 2003.

[109] P. Burt and R. Kolczynski, "Enhanced image capture through fusion," in *Computer Vision, 1993. Proceedings., Fourth International Conference*, pp. 173–182, IEEE, 1993.

[110] G. Piella and H. Heijmans, "A new quality metric for image fusion," in *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference*, vol. 3, pp. III–173, IEEE, 2003.

[111] J. Maintz and M. Viergever, "An overview of medical image registration methods," *UU-CS*, no. 1998-22, 1998.

[112] B. Zitova and J. Flusser, "Image registration methods: a survey," *Image and vision computing*, vol. 21, no. 11, pp. 977–1000, 2003.

[113] G. Stockman, S. Kopstein, and S. Benett, "Matching images to models for registration and object detection via clustering," *Pattern Analysis and Machine Intelligence, IEEE Transactions*, no. 3, pp. 229–241, 1982.

[114] S. Richard, "Image alignment and stitching: A tutorial," *Microsoft Research, September*, vol. 27, 2004.

[115] A. Rangarajan, H. Chui, and J. Duncan, "Rigid point feature registration using mutual information," *Medical Image Analysis*, vol. 3, no. 4, pp. 425–440, 1999.

[116] C. Schutz, T. Jost, and H. Hugli, "Multi-feature matching algorithm for free-form 3d surface registration," in *Pattern Recognition, 1998. Proceedings. Fourteenth International Conference*, vol. 2, pp. 982–984, IEEE, 1998.

Glossary

List of Acronyms

rgb	red, green and blue
cmyk	cyan, magenta, yellow and black
xrf	X-ray fluorescence
ir	infrared
uv	ultraviolet
lcd	Liquid Crystal Display
lcos	Liquid Crystal on Silicon
dlp	Digital Light Processing
dmd	Digital Micromirror Device
ccd	Charge-Coupled Device
cmos	complementary metal oxide semiconductor
crt	Cathode Ray Tube
orm	Object Recognition Memory
cns	Central Nervous System
dnr	dynamic range
dft	discrete Fourier transform
hdr	high dynamic range
sfr	spatial frequency response
mi	mutual information

dof	depth of field
icc	International Color Consortium
cie	Commission Internationale de l'Éclairage
sps	Single Phase Shifting
psp	Phase Shifting Profilometry
mps	Multiple Phase Shifting
csl	Coded Structured Light
lab	CIE-L*a*b*
eci	European Color Initiative
psf	point spread function
snr	signal-to-noise ratio
tof	Time-of-Flight
sad	sum of absolute differences
ssd	sum of squared differences
api	application programming interface
sift	scale-invariant feature transform
lidar	Light Detection and Ranging
fadgi	Federal Agencies Digitization Initiative Still Image Working Group
mems	Micromechnical System
pcb	Printed Circuit Board
smd	Surface Mount Device
usb	Universal Serial Bus
dc	Direct Current
dof	Degrees of Freedom
dslr	Digital Single Lens Reflex
gpu	Graphical Processing Unit
gui	Graphical User Interface
mtf	Modulation Transfer Function
lut	look-up table

rms	root mean square
icp	iterative closest point