Light Field Applications to 3-dimensional Surface Imaging

by

Wenxian Hong

Submitted to the Department of Mechanical Engineering in partial fulfillment of the requirements for the degree of Bachelor of Science in Mechanical Engineering at the MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2009

© Massachusetts Institute of Technology 2009. All rights reserved.

Author ..........................................................
Department of Mechanical Engineering
August 7, 2009

Certified by ..................................................
Douglas P. Hart
Professor of Mechanical Engineering
Thesis Supervisor

Accepted by ..................................................
John H. Lienhard V
Collins Professor of Mechanical Engineering
Chairman, Undergraduate Thesis Committee
Light Field Applications to 3-dimensional Surface Imaging

by

Wenxian Hong

Submitted to the Department of Mechanical Engineering on August 7, 2009, in partial fulfillment of the requirements for the degree of Bachelor of Science in Mechanical Engineering

Abstract

The structure of light around a scene may be contained in a 4-dimensional array known as a light field. This thesis describes methods for acquiring and manipulating light fields for applications in 3-dimensional imaging. By actively sampling parts of the wavefront impinging on a lens, or using microlens arrays and patterned sinusoidal masks to modulate the rays reaching a camera, both the spatial distribution and directionality of light may be captured to produce light fields. Simple depth estimation algorithms using stereo and focus measures are then applied to recover quantitative depth information. Experiments on real-world light fields demonstrate their utility in performing digital refocusing, reconstructing occluded objects as well as accurately estimating depth and shape. The performance of the algorithms developed are discussed theoretically and compared empirically.

Thesis Supervisor: Douglas P. Hart
Title: Professor of Mechanical Engineering
Acknowledgments

I would like to thank my advisor, Prof. Douglas Hart, for his support and guidance throughout my research. I am also very grateful to Dr. Ankit Mohan, Prof. Ramesh Raskar and the Camera Culture group at the MIT Media Lab for their helpful insights and for generously letting me use their equipment. Finally, I would like to acknowledge Dr. Amit Agrawal for providing the heterodyne camera datasets.

On a personal note, thanks to my parents and friends for their constant encouragement, and to my professors and teachers for the enriching experience that is MIT.
Contents

1 Introduction ........................................ 6

2 Light Field Basics .................................. 9
   2.1 The Plenoptic Function ....................... 9
   2.2 The 4D Light Field ......................... 10
   2.3 Light Field Parameterization ............. 11
   2.4 Uses of Light Fields ....................... 12

3 Active Wavefront Sampling ...................... 13
   3.1 Depth From Defocus (DFD) ................. 13
   3.2 Active Wavefront Sampling ............... 16
   3.3 Advantages of AWS ....................... 18
   3.4 Implementation of AWS ................... 18
   3.5 Image Processing ......................... 19
   3.6 Relation to Light Fields ................. 20

4 Plenoptic Cameras .................................. 22
   4.1 Dense Camera Arrays ...................... 22
   4.2 Plenoptic Cameras ......................... 23

5 Optical Heterodyning ............................. 25
   5.1 Introduction to Heterodyning ............ 25
   5.2 Theory of Mask-based Heterodyning ..... 26
      5.2.1 Parametrization .................... 26
### 5.2.2 Generic Light Field Modulation

### 5.2.3 Optimal Light Field Modulation Using Sinusoidal Mask

### 5.2.4 Obtaining Light Field Through Demodulation

### 5.2.5 Extension to 4D Light Field Capture

### 5.2.6 Recovering Full-resolution Image

### 5.3 Practical Implementation

#### 5.3.1 Practical Considerations for Mask

#### 5.3.2 Design of Heterodyne Camera

### 5.4 Advantages and Limitations

### 6 Depth Reconstruction

#### 6.1 Overview of Depth Estimation Techniques

#### 6.2 Synthetic Aperture Focusing

##### 6.2.1 Theoretical Analysis of Refocusing

##### 6.2.2 Calibrating Parallax for Planar Camera Arrays

#### 6.3 Experimental Methodology

##### 6.3.1 Data Acquisition

##### 6.3.2 Description of Focus and Stereo Algorithms

#### 6.4 Results

##### 6.4.1 Digital Refocusing

##### 6.4.2 Depth Estimation

##### 6.4.3 3D Surface Imaging

#### 6.5 Discussion

##### 6.5.1 Comparison of Depth-from-Focus and Depth-from-Stereo

##### 6.5.2 Performance Potential of Algorithms

### 7 Conclusions

#### 7.1 Future Directions
Chapter 1

Introduction

The structure of light that permeates a scene contains a rich amount of information that may be used to characterize its geometry. Unfortunately, conventional cameras only record a two-dimensional projection of our 3D world and thus sacrifice all depth information. Even the human eye has to rely on a variety of subtle cues like texture variations, color and shading in order to perceive depth from this two-dimensional projection. Our perception of reality is so strongly shaped by these 3D depth cues that we unconsciously and seamlessly integrate them into the way we communicate and respond to our environment.

For this reason, moving away from a 2D portrayal of the world to a 3D one is not only desirable, but a natural progression. Indeed, the potential for such a technology is tremendous. Existing 3D capture devices such as laser scanners and time-of-flight rangefinders have been used in applications as diverse as defect detection, robotic navigation, sports and CGI-based entertainment. Better 3D display technology is also on the horizon, which will change how we interact with computers and other people.

Presently, however, most work on visual 3D reconstruction has focused on binocular stereo systems. By capturing two different viewpoints and mimicking the way our eyes interpret visual stimuli, we may generate 3D representations of the scene. However, binocular systems face the drawback of requiring two separate cameras, which increases costs and complicates image processing.
Ideally, we would like to capture as much of the structure of light in space as we can by using as simple an optical arrangement as possible, for instance by using a single lens or requiring only a single snapshot. Presumably, after capturing all the information available in the light received, we may not only infer depth, but also render novel viewpoints, look behind occlusions and manipulate light in a myriad of other meaningful possibilities.

Prior to capture, we first require a theoretical formulation of the information content of light at every point in space. We shall see that such a formulation may be achieved by considering a 4D array known as the light field. Simply put, the light field describes the amount of light along all the rays that intersect a scene of interest. Equipped with this knowledge, we may then model a real-world scene without explicit knowledge of its geometry, enabling us to move towards a 3D or even higher-dimensional representation.

Furthermore, because the light field contains a wealth of information about the scene in a computationally-friendly format, it is an ideal candidate for bridging real-world geometry and a computer’s analysis of the scene. Simple algorithms may then be applied to the light field to extract a range of desired information.

In particular, this thesis focuses on the role of light fields in depth reconstruction and 3D surface imaging. The general approach is to understand, acquire, and then utilize these light field arrays, as outlined below:

First, the light field structure is described in detail in Chapter 2, together with the necessary mathematical framework.

Chapters 3-5 then present various methods for acquiring light fields. Chapter 3 focuses on Active Wavefront Sampling (AWS), a technique that works by taking multiple samples of the optical wavefront impinging on a lens to calculate the depth of scene features. Although AWS was specifically devised for 3D-surface imaging, this approach may be also understood in the context of light fields. Chapter 4 introduces the plenoptic camera, which captures multiple viewpoints of a scene in a single snapshot by using a microlens array. Chapter 5 deals with a non-refractive counterpart to the plenoptic camera using mask-based optical heterodyning. By inserting a pat-
terned sinusoidal mask into a conventional camera, the light rays entering the camera may be modulated so that their directions can also be captured by the 2D imaging sensor.

In Chapter 6, we examine the specific application of light fields to depth reconstruction. We develop algorithms to aggregate the information within the arrays so as to extract useful information about the scene geometry.

Finally, Chapter 7 provides a summary and indicates directions for future work.
Chapter 2

Light Field Basics

The light around us may be understood in terms of rays with associated spatial, angular and spectral distribution. A complete description of the structure of ambient light must thus account for these properties. This chapter introduces the plenoptic function, which describes the set of all light rays passing through space-time. We pay particular attention to a simplified but more useful subset, known as the light field.

2.1 The Plenoptic Function

The full visual appearance of a scene can be modeled as a single function, called the plenoptic function, where the word ‘plenoptic’ stems from the Latin roots for ‘complete’ and ‘view’ [1]. In essence, the plenoptic function records the amount of light along each ray (formally measured by the radiance $L$) anchored at any location in space $(x, y, z)$ for every orientation $(\theta, \phi)$ at any time $t$, for a given wavelength $\lambda$. It can therefore be thought of as describing a 7D ray space, $L(x, y, z, \theta, \phi, \lambda, t)$.

From a practical standpoint, however, the spectral dimension $\lambda$ is often encoded in the measurement of radiance, as both the human eye and modern CCD sensors sample light intensity at the wavelengths of red, green and blue light. Furthermore, for a static region illuminated by an unchanging arrangement of lights, the temporal dimension $t$ is unnecessary. As such, rays in space are typically parameterized by their three spatial coordinates $(x, y, z)$ and two angles ($\theta$ and $\phi$) alone, thus giving a
Since the plenoptic function captures all visible information of a scene, the process of visual perception involves the extraction of appropriate subsets. A photograph, for instance, measures the plenoptic function over a finite range of directions for a fixed camera position. Because only two of the parameters \((\theta, \phi)\) may vary, a photograph is a 2D slice of the 5D plenoptic function.

![Diagram of a ray in 3D space parameterized by position \((x, y, z)\) and direction \((\theta, \phi)\).[16]](image)

**2.2 The 4D Light Field**

While the plenoptic function gives a comprehensive description of any scene, it has an undesirable redundancy. By noting that the radiance along a ray remains constant unless blocked, the dimensionality may actually be reduced by one. Since light rays propagating in the same direction are effectively similar, it suffices to represent them at a single point rather than at all points along the line. Therefore, the 5D plenoptic function may be simplified to a 4D subset, called the light field.

Formally, the 4D light field is defined as the radiance along all rays in free space [16]. The free space assumption restricts the scenes which may be represented to those without occlusions, attenuating medium (e.g. fog) and phase-altering elements (e.g. phase masks). Nonetheless, these limitations are usually not of concern as most...
scenes exist in a clear medium (e.g. air) and are contained within some bounded area, like the convex hull of an object, into which the camera does not enter.

2.3 Light Field Parameterization

The 4D light field can be parameterized in a number of ways. The parametrization that follows most naturally from that of the plenoptic function represents each ray in terms of its two angles of propagation ($\theta, \phi$), and its position on the reference ($x, y$) plane (Figure 2-2(a)). Note that only two position coordinates are needed instead of three, because of the constancy of each ray along its direction of propagation. This spherical-cartesian representation allows light rays travelling in all directions to be represented with equal resolution.

![Figure 2-2: (a) Spherical-Cartesian parameterization of light field. (b) Two-Plane parameterization (2PP) of light field. [19]](image)

More commonly, the light field is parameterized using the two-plane parameterization (2PP), first proposed in [2] (Figure 2-2(b)). In the 2PP, a light ray is defined by its points of intersection with two reference planes, ($u, v$) and ($s, t$), where both planes are arbitrarily defined with some positive separation $d$. Such a representation is also known as a light slab. Though the reference planes are assumed to be continuous and infinite, in practice they are discretely sampled over a finite region to give the value of a color triplet ($r, g, b$). Intuitively, one may think of a two-plane light
field \( L(u, v, s, t) \) as a set of perspective images of the u-v plane, taken from multiple viewpoints on the s-t plane.

The 2PP representation has several important advantages. For one, it allows for greater computational efficiency, because mapping an image pixel to its \((u, v, s, t)\) coordinates only requires a simple projective map, compared to more complicated transforms using spherical or cylindrical coordinates. Therefore, though the 2PP inherently cannot represent all rays, such as those travelling parallel to either plane, it is often still the preferred parameterization. Multiple light slabs may even be used to provide wider coverage and more uniform sampling of rays [2]. Finally, by placing one of the reference planes at infinity, the 2PP conveniently becomes the spherical-cartesian representation.

In this thesis, we will primarily employ the two-plane parametrization.

2.4 Uses of Light Fields

The rich information content and ease of manipulation of light fields make them highly useful for many applications in computer graphics and vision. Light fields have been used extensively as a means of quickly rendering novel views of a scene by extracting appropriate 2D slices of the 4D function [2]. Since the light field contains all the rays that intersect a scene, we can easily generate new synthetic camera views on the fly by resampling the acquired rays, without being limited by the geometric complexity of the scene. Other uses of light fields include synthetic aperture photography [3], reconstructing views around occlusions [6], digital refocusing [4] and glare reduction.

In particular, this thesis focuses on the use of light fields to estimate the shapes of scenes. Presumably, the geometry of a scene can be found by determining the depth, and thus 3D position, at every visible point. Chapter 6 will discuss techniques that utilize the information from light field models of real-world scenes to construct depth maps and render 3D surfaces.
Chapter 3

Active Wavefront Sampling

Having briefly formalized the concept of light fields, the next three chapters will describe various methods to acquire them. Before doing so, it is instructive to examine the optical information already available across the camera lens aperture. This chapter will then introduce the technique of Active Wavefront Sampling (AWS), which captures part of the structure of light impinging on different subregions of the lens.

3.1 Depth From Defocus (DFD)

It is well known that a single image captured with a standard monocular lens contains some depth information in the form of defocus blur diameter. Consider the optical system in Figure 3-1(a). The sensor at distance $Z_{\text{cco}}$ behind the lens brings a point object at distance $Z_{fp}$ in front of the lens into focus. In this case, the object is located on the in-focus plane and there is no defocus blur. If the object is placed nearer or farther from the lens as in Figure 3-1(b) and 3-1(c), the object is now out-of-focus and its image is a blur-circle. The diameter of the defocus blur increases as the object moves away from the in-focus plane in either direction.

Specifically, the relationship between the blur diameter $d$ and object depth $Z_o$ may be found from similar triangles and the lens imaging condition, giving the following...
Figure 3-1: Principle of Depth from Defocus: In-focus point object (a) forms a point image. Near object (b) and far object (c) form blur circles of diameter $d$. 
expression:

\[ \frac{d}{D} = 1 - \frac{\frac{1}{f} - \frac{1}{Z_o}}{\frac{1}{f} - \frac{1}{Z_{fp}}} , \]  

(3.1)

where \( f \) is the focal length of the lens and \( D \) is the diameter of the lens aperture [18].

For simplicity, we invoke the paraxial approximation so that the system is invariant to transverse shifts. Noting that the denominator in Equation 3.1 is the reciprocal of the lens-sensor distance \( Z_{cdd} \), the expression may also be rewritten as: [18]

\[ \frac{d}{D} = Z_{cdd} \left( \frac{1}{Z_o} - \frac{1}{Z_{fp}} \right) \]

(3.2)

A normalized plot of Equation 3.2, shown in Figure 3-2, reveals two interesting features. First, the sensitivity of depth-from-defocus methods is greater for objects between the lens and the in-focus plane, as seen from the steeper gradient of the curve. Second, for a given diameter measurement \( d \), there are two possible object positions: one between the lens and in-focus plane \( (\frac{Z_o}{Z_{fp}} < 1) \), and the other beyond the in-focus plane \( (\frac{Z_o}{Z_{fp}} > 1) \). Indeed, from Figure 3.1 we can intuitively see that objects at two different depths may give rise to the same blur diameter. Without some prior knowledge of scene geometry, it is not possible to resolve this depth ambiguity.

![Figure 3-2: Plot of normalized blur spot diameter against normalized object distance](image-url)
The main idea behind DFD methods is then to accurately determine the blur diameter of each object feature so as to calculate its depth. Even so, this task is not always easy. Feature-rich scenes may give rise to overlapping blur spots, which could be difficult to distinguish. The presence of occlusions and noise in the scene also produce less-than-ideal Point Spread Functions (PSF), which hinder the estimation of blur diameter and reduce the robustness of DFD algorithms [9]. Finally, the problem of depth ambiguity remains.

3.2 Active Wavefront Sampling

Suppose instead that the entire lens is blocked except for an eccentric pinhole. In effect, this modified aperture selectively samples the entire wavefront reaching the lens by allowing only one ray to pass through. For an object situated on the in-focus plane, the image would appear sharp and stationary (Figure 3-3(a)). But if the object is near, as in Figure 3-3(b), the eccentric aperture causes the image to be displaced upwards. Conversely, the image of a far object is displaced downwards (Figure 3-3(c)). In fact, as the aperture is rotated around a circle centered on the optical axis of the lens, the image on the sensor will also be seen to rotate in a circle, the diameter of which is proportional to the extent of blur. Images of objects located between the in-focus plane and the lens will rotate 180° out of phase with those located beyond the in-focus plane. Together, the diameter and direction of rotation unambiguously give the depth of the object according to Equation 3.2.

One may therefore take a series of images in which the aperture is moved around and perform displacement analysis on the images to determine target feature depths. The aperture need not be rotated in a circle, but as long as its exact motion is known, in theory the object distance may be recovered. Adelson and Wang termed such a method “single-lens stereo” [1]; more recently, this principle has formed the basis for a 3D imaging technique known as Active Wavefront Sampling [18].
Figure 3-3: Principle of Active Wavefront Sampling: In-focus point object (a) remains stationary, but the images of near objects (b) and far objects (c) are displaced upwards and downwards respectively.
3.3 Advantages of AWS

The observation that two images from different wavefront sampling positions are slightly displaced from each other leads to a natural comparison with traditional stereo methods. At first glance, since the effective stereo baseline of an AWS system is limited to the size of the lens aperture, it might seem that the depth resolution of AWS is inherently poorer than stereo systems with larger baselines.

However, the AWS system has several advantages. One significant benefit is that AWS only requires a single optical path. This not only reduces the cost of optics, but also allows AWS to be used in applications where 3D information is desired but a large baseline is unsuitable, such as endoscopy and microscopy. Second, due to the smaller baseline, AWS minimizes the classic stereo problem of correspondence and requires a smaller search space to find matching points. Furthermore, the high sampling rate in AWS allows many more image points to contribute to the depth estimation, thereby improving accuracy and making AWS more robust in the presence of slight occlusions. Real-time processing is even achievable by decreasing the number of AWS sampling positions to as few as two (like stereo).

3.4 Implementation of AWS

Frigerio implemented a simple mechanical version of AWS using a rotating aperture as described earlier [18]. This design was chosen for its ease of construction and computational convenience (given a known circular sampling pattern). A bearing-supported rotating disk with an eccentric aperture is placed directly next to the lens and actuated by a stepper motor via a belt train (Figure 3-4). The system makes use of a 50mm Nikorr lens and 8-bit digital camera, giving a maximum frame rate of 30Hz. Finally, a sinusoidal pattern is projected onto the scene to increase the signal-to-noise ratio.

The AWS system may also be implemented with electronic components. Instead of using a rotating aperture, one may place an LCD device at the lens. By turning
Figure 3-4: Photograph of mechanical AWS device built in [18]. The camera lens (mounted over the module) is removed to reveal the internal components.

parts of the LCD on or off, different subregions of the lens may be sampled. This design avoids the need for moving parts, but image quality potentially suffers due to light attenuation. Alternatively, one may also use a Digital Micromirror Device (DMD) in place of the rotating aperture. The DMD is a chip that contains over a million tiny mirrors, whose tilt may be individually controlled using a computer. When activated, each mirror reflects light from a different part of the lens onto the sensor, thereby achieving active wavefront sampling (Figure 3-5).

3.5 Image Processing

After the sequence of images taken at various sampling positions is obtained, image processing algorithms may be applied in order to measure depth. For the mechanical implementation of AWS, Frigerio adopts a coarse-to-fine multi-image approach [18]. First, a sparse array image correlation algorithm (a variant of block-matching) is applied to successive images to determine their pixel-wise displacement relative to a reference image. A gradient-based optical flow algorithm is then employed to find the remaining sub-pixel displacement. Frigerio also discusses methods to compensate for illumination variation in the scene and the effect of lens aberrations, so as to increase robustness. Admittedly, most of these techniques are specific to the case of a
pure aperture rotation; other implementations require the algorithms to be modified accordingly. Nonetheless, the ultimate goal of image processing remains the same: to accurately determine the image displacement for every feature in the scene as the wavefront hitting the lens is actively sampled.

### 3.6 Relation to Light Fields

In effect, the data collected from AWS may be thought of as a light field inside the camera. Under the two-plane parametrization, the main lens is chosen to be the $u$-$v$ plane, while the sensor forms the other $s$-$t$ plane. As the lens aperture is rotated, the sensor captures light field views from different $(u, v)$ coordinates on the lens plane. However, the light field obtained from a pure aperture rotation is not truly 4-dimensional because the $(u, v)$ sampling positions are correlated by the radius of the eccentric aperture, giving variation in only one (angular) dimension. To obtain the two-dimensionality of the lens plane, images must be taken at various sampling radii. Electronic implementations of AWS readily achieve this as they are not restricted to...
a circular sampling pattern, and may conceivably yield light fields with much higher spatial and directional resolution.

The relation between AWS and light fields suggests that some of the techniques used to determine depth from light fields may also be extended to AWS, on top of the existing methods described in [18]. These include depth-from-focus, depth-from-stereo and gradient-based methods, which will be discussed in Chapter 6. Furthermore, AWS datasets may theoretically be used for synthetic aperture photography, digital refocusing, depth-of-field extension, and other applications of light fields.
Chapter 4

Plenoptic Cameras

The disadvantage of Active Wavefront Sampling is that multiple images are required for a complete depth estimation. Instead, it might be preferable to acquire all the information from the scene with a single snapshot. In this chapter, we briefly describe some existing devices used to acquire light fields, namely camera arrays and plenoptic cameras.

4.1 Dense Camera Arrays

Perhaps the most straightforward way to capture an entire light field at once is to use a dense planar array of cameras. Each camera occupies a different \((s, t)\) coordinate and captures a different viewpoint of the scene. Projecting the image pixels to the \((u, v)\) plane requires only a simple calibration procedure and keystone correction. Multi-camera arrangements like the one constructed at Stanford (Figure 4-1(a)) not only allow for real-time processing of dynamic scenes, but can also see through partially occluded environments like foliage due to their larger effective aperture [6]. Recently, ViewPLUS Inc. has also developed a compact 5x5 camera array system called the ProFUSION 25, which easily interfaces with a computer (Figure 4-1(b)). We have used the ProFUSION 25 to acquire several light field datasets found in Chapter 6.
Figure 4-1: (a) Stanford Multi-Camera Array, consisting of 8x16 synchronized CMOS cameras (b) Compact ProFUSION 25 camera array used to capture some of the datasets in Chapter 6.

4.2 Plenoptic Cameras

In a conventional camera, each sensor pixel simply sums the incident cone of light rays regardless of direction, effectively only capturing a 2D projection of the 4D light field (Figure 4-2(a)). On the other hand, plenoptic cameras retain directional information and allow the 2D sensor to sense 4D radiance coming from different parts of the lens aperture.

Pioneered by Adelson and Wang [1], the design of plenoptic cameras involves inserting a microlens array between the main lens and the sensor, as shown in Figure 4-2(b). Rays converging onto each microlens are directed to different areas of the underlying macropixel, according to their incident direction. Hence, each sensor pixel is associated with a ray passing through the \((s, t)\) microlens and a \((u, v)\) point on the main lens, producing a 4D light field array.

Furthermore, the subimage formed within each macropixel represents the view of the main lens aperture \((u-v)\) as seen from a different \((s, t)\) microlens position. Extracting the same pixel in each micro-image thus amounts to gathering all the rays coming from a particular \((u, v)\) subaperture of the main lens and rearranging to form the image that would have resulted if only that subaperture were exposed on the lens. If each microlens covers \(k\) pixels in one dimension, one then captures \(k\) viewpoints of
Figure 4-2: (a) Image capture in a conventional camera. The sensor sums the incident bundle of light, thus losing the lighting directionality. (b) Schematic of plenoptic camera in [1, 4]. The microlens array separates the incoming rays into subimages within macropixels and may be modeled as pinholes. The dotted lines show how each sensor pixel is associated with a \((u, v, s, t)\) coordinate.

The scene, but the spatial resolution is also reduced by a factor of \(k\) in that dimension. Using these reduced subaperture views, Adelson and Wang were able to reconstruct the depth of photographed objects.

More recently, Ng et al. created a portable hand-held version of the plenoptic camera [4]. By rearranging the light rays (or equivalently, rearranging the pixels) such that they converge on a different synthetic focal plane, Ng et al. show that such a plenoptic camera can achieve post-capture digital refocusing.
Chapter 5

Optical Heterodyning

While plenoptic cameras can capture multiple viewpoints of a scene, their use of refractive elements renders makes susceptible to the effects of spherical/chromatic aberrations, coma and misalignment. In contrast, Veeraraghavan et al. describe a heterodyne light field camera that uses only a non-refractive sinusoidal attenuating mask to modulate and recover the 4D light field [10]. Unlike other plenoptic cameras which sample individual rays arriving at the sensor, the heterodyne camera senses linear combinations of rays in the Fourier domain. This technique of optical heterodyning using non-refractive modulators will be the subject of this chapter.

5.1 Introduction to Heterodyning

Heterodyning is widely used in signal processing and telecommunications. The basic idea is to modulate a baseband signal with a higher frequency carrier so as to band-shift the signal into an intermediate passband for efficient transmission. At the receiver end, the incoming signal is then demodulated to recover the original signal.

This mixing of frequencies to generate new frequencies follows directly from the Modulation Theorem in the Fourier domain [17], where multiplying a baseband signal $s(x)$ by a cosine waveform of frequency $f_0$ produces two spectral copies of the signal:

$$
\mathcal{F} [\cos (2\pi f_0 x) s(x)] (f_x) = \frac{1}{2} (F(f_x - f_0) + F(f_x + f_0)) ,
$$

(5.1)
where \( \mathcal{F}[s(x)](f_x) = F(f_x) \) is the Fourier transform of \( s(x) \).

The effect of introducing a sinusoidal light field modulator is very similar to radio heterodyning. To sense a 4D light field using only a 2D sensor, the light field must be modulated so that the angular information may also be captured. Placing a sinusoidal mask between the lens and sensor achieves this by creating spectral tiles of the light field in the 4D frequency domain. In this way, the angular dimensions of the light field that are usually lost are now modulated to higher spatial frequencies, which the 2D sensor can then detect. Recovering the light field is analogous to demodulation, and involves redistributing the high spatial frequencies of the 2D sensed signal back to their original lower angular dimensions in the 4D light field space.

### 5.2 Theory of Mask-based Heterodyning

In this section we describe the theory by which an appropriately designed and positioned sinusoidal mask achieves the required modulation for light field heterodyning. We refer extensively to the analysis done by Veeraraghavan et al. in [10], [11] and [12].

#### 5.2.1 Parametrization

We begin by parametrizing the light field entering the camera using the standard two-plane parametrization, described in Section 2.3. Consistent with the notation in [10], the sensor plane is defined to be the \( x-y \) plane and the lens aperture forms the other \( \theta-\phi \) plane, separated by distance \( v \). Note however that while the \( \theta-\phi \) lens plane represents the angular spectrum of the light field, it remains parameterized by spatial coordinates, so that an increase in \( (\theta, \phi) \) actually refers to an in-plane translation rather than a change in angle.

For simplicity, we first consider a 2D flatland scenario with only \( x \) and \( \theta \) dimensions. In this case, a light field \( l(x, \theta) \) passes through a 1D mask pattern \( c(y) \), as shown in Figure 5-1. The extension to the general 4D case is later shown to be straightforward.
Figure 5-1: Schematic layout of lens, mask and sensor in a heterodyne light field camera. In flatland, a 1D mask $c(y)$ is placed in front of a 1D sensor to modulate a 2D light field. The lens and sensor are defined to be the $\theta$- and $x$- planes respectively.

As a further approximation, the light field is assumed to be rectangularly band-limited to $(fo, feo)$, i.e. in the Fourier light field space, $L(f_x, f_\theta) = 0 \forall |f_x| \geq f_{x0}, |f_\theta| \geq f_{\theta0}$ (see Figure 5-2(a)). This assumption allows us to neglect the effects of aliasing (discussed in Section 5.4), and is mostly valid given that the energy in the high frequency band is usually small for natural scenes [12].

5.2.2 Generic Light Field Modulation

First, we consider light field modulation by a generic non-refractive mask. A 2D light field that is attenuated by a 1D mask pattern $c(y)$ is in fact modulated by a 2D Ray Modulation Function (RMF), $m(x, \theta)$. To see this, we observe the dependence of $y$ on $x$ and $\theta$ from Figure 5-1. The RMF for a mask placed at distance $d$ from the sensor is then given by

$$m(x, \theta) = c(y) = c\left(\frac{d}{v}\theta + (1 - \frac{d}{v})x\right)$$  \hspace{1cm} (5.2)

From Equation 5.2, it is clear that the position of the mask changes how it modulates the incoming light field. If the mask is placed at the sensor (i.e. $d = 0$), then all rays converging to a particular point on the sensor are attenuated equally regardless
of \( \theta \). The RMF is then \( m(x, \theta) = c(y = x) \). Conversely, if the mask is placed at the lens aperture (i.e. \( d = v \)), then rays arriving at the same subregion of the lens are affected in the same manner regardless of their eventual direction, and thus the RMF only depends on \( \theta \): \( m(x, \theta) = c(y = \theta) \).

In the 2D Fourier domain, Equation 5.2 implies that the RMF of a 1D mask \( M(f_x, f_\theta) \) lies entirely on a line whose slope depends on the mask placement (Figure 5.2.2). Specifically, the slope angle \( \alpha \) with respect to the \( f_x \) axis is given by

\[
\alpha = \arctan \left( \frac{d}{v - d} \right)
\]

Indeed, by taking the Fourier transform of Equation 5.2, the mask RMF may be shown to be

\[
M(f_x, f_\theta) = \mu^2 C \left( \frac{1}{f_x^2 + f_\theta^2} \right) \delta(f_x \cos \alpha - f_x \sin \alpha), \tag{5.4}
\]

where \( C \) denotes the Fourier transform of the 1D mask and \( \mu = \frac{1}{\sqrt{(\frac{d}{v})^2 + (1-\frac{d}{v})^2}} \) [13]. Due to the \( \delta \)-function term, the mask RMF is concentrated along a line in Fourier space given by \( f_x \cos \alpha - f_x \sin \alpha = 0 \), as previously noted.

As the generic mask only attenuates light rays without bending them, the modulated light field \( l_m(x, \theta) \) that leaves the mask is simply the product of the original light field and mask RMF [11], viz.

\[
l_m(x, \theta) = l(x, \theta) m(x, \theta) \tag{5.5}
\]

Since multiplication in the primal domain corresponds to convolution in the Fourier domain, Equation 5.5 may also be written as

\[
L_M(f_x, f_\theta) = L(f_x, f_\theta) \otimes M(f_x, f_\theta), \tag{5.6}
\]

where \( \otimes \) denotes the convolution operator, and the uppercase variables denote the respective Fourier domain representations [11].
As such, an incoming light field is modulated by convolving with a mask RMF lying along a line in 2D Fourier space. Depending on the exact mask spectrum and mask position, the shape of the modulated light field can take various forms. For instance, if the mask RMF consists of an infinite series of impulses along a slanted line, the resultant modulated light field also comprises infinitely many spectral replicas of the light field along the slanted line centered around each impulse, as shown in Figure 5-2(a). Such a mask RMF corresponds to a pinhole array placed close to the sensor, analogous to the microlens camera design described in Section 4.2 [11]. With these spectral copies, the sensed image, which corresponds to the horizontal slice of the modulated light field (in green) [5], now captures some of the angular information in the light field in the form of higher spatial frequencies. However, as with any pinhole array, there is significant loss of light and poor signal-to-noise ratio, as evidenced by the sensor slice capturing only a limited portion of the continuum of spectral copies. Moreover, the overlap of spectral tiles is not ideal and leads to aliasing.

5.2.3 Optimal Light Field Modulation Using Sinusoidal Mask

The task that remains is to optimize the mask pattern and mask position so as to efficiently capture all the information in the light field with minimal loss of energy and no wastage of sensor resolution. To this end, several impulses spaced apart correctly and aligned at an appropriate angle will be sufficient for modulation, rather than an infinite train of impulses. In the primal domain, this optimal mask RMF corresponds to a sum of several cosines along with a DC term. The result of convolution (Equation 5.6) is then an optimally modulated light field, as shown in Figure 5-3.

5.2.3.1 Optimal Mask Position

Consider the light field bandlimited to \((f_{x0}, f_{\theta0})\) and discretized by a desired frequency resolution \(f_{\theta R}\) along the \(f_{\theta}\) axis, shown in Figure 5-3. For optimal sampling along the horizontal \(f_x\) slice, the spectral tiles should line up alongside without overlapping and be vertically displaced from one another by \(f_{\theta R}\). Hence, the optimal slant angle
Figure 5-2: (a) Light field modulation using a generic mask. In 2D Fourier space, a bandlimited light field spectrum (left) convolves with a mask modulation function (center) to produce spectral replicas along a tilted line (right). The slope angle $\alpha$ depends on the position of this mask, as per Eq. 5.3. The sensor image is a horizontal slice of the light field. Prior to modulation, the sensor cannot detect angular variation. However, after modulation, the sensed image (green box) now captures some information in the angular dimensions. (b) In this case, the mask RMF comprising an infinite train of impulses corresponds to an array of pinholes placed between the lens and sensor. Also, the overlapping spectral tiles indicate overlapping sub-images in the primal domain.
Figure 5-3: Optimal modulation of rectangularly bandlimited light fields (taken from [10]). (left) The optimal mask RMF consists of \((2p + 1)\) impulses separated by fundamental frequency \(f_0\) and tilted at angle \(\alpha_{opt}\). (right) After convolution, spectral slices of varying \(f_0\) content now appear along the horizontal \(f_x\) axis, allowing the sensed image (red box) to capture the entire 2D light field spectrum. The absence of gaps and extraneous spectral copies minimizes light loss and wastage of sensor pixels.

\[\alpha_{opt} \text{ is given by} \]

\[\alpha_{opt} = \arctan \left( \frac{f_{0R}}{2f_{x0}} \right) \quad (5.7)\]

With the optimal slant angle \(\alpha_{opt}\) in hand, we can then compute the optimal mask position \(d\) via Equation 5.3 for a given lens-sensor distance \(v\). Practically speaking, the spatial bandwidth of an incoming lightfield typically exceeds the angular resolution (which is inversely related to aperture size). Hence, \(\alpha\) is often small (4-5°) and the mask must thus be placed closed to the sensor.

5.2.3.2 Optimal Mask Pattern

As shown above, optimal light field modulation requires a mask RMF \(M(f_x, f_\theta)\) consisting of several impulses along a line in 2D Fourier space. From Equation 5.4, it follows that the Fourier transform of the mask code \(C(f_\theta)\) should also be a set of impulses. Design of an optimal mask pattern then involves determining the appropriate spacing of impulses and the optimal number of impulses.

If the light field is rectangularly bandlimited, as in Figure 5-3, the \((2p+1)\) impulses of the mask code should be evenly separated by some fundamental frequency \(f_0\). In
mathematical terms, the Fourier transform of the mask code \( c(y) \) may be written as

\[
C(f_y) = \sum_{i=-p}^{p} \delta(f_y - if_0)
\]  

(5.8)

The fundamental frequency \( f_0 \) can be obtained from Equation 5.4 by substituting \( f_x = 2f_{x0} \) and \( f_\theta = f_{\theta R} \), viz.

\[
f_0 = \mu \sqrt{4f_{x0}^2 + f_{\theta R}^2}
\]  

(5.9)

Furthermore, the optimal number of impulses depends on the assumed angular bandwidth of the lightfield \( f_{\theta 0} \) and desired frequency resolution \( f_{\theta R} \). Excessive numbers of impulses create too many spectral copies which exacerbates light loss without detecting any additional information. Conversely, too few impulses will fail to capture all the available information. From Figure 5-3, we see that the number of impulses \((2p + 1)\) should equal to the total number of angular samples, i.e.

\[
2p + 1 = \frac{2f_\theta}{f_{\theta R}}
\]  

(5.10)

The resultant optimized mask is a sum of \( p \) harmonic cosines of fundamental frequency \( f_0 \) and a DC term. With these optimal choices of \( \alpha \), \( f_0 \) and \( p \), the modulated light field does not contain any extraneous spectral copies or gaps within the horizontal slice. Hence, the 1D sensor now captures all the information in the original light field with minimal light loss and optimal sensor pixel usage.

If we relax the rectangular bandlimited assumption and consider differently shaped light field spectra, the optimal mask RMF should intuitively consist of unequally spaced impulses, so that the spectral copies arising from modulation may tessellate with each other along the horizontal slice (Figure 5-4) [11]. The resulting mask is then a sum of non-harmonic cosines. Presumably, given prior information on the shape of an incoming light field spectrum, we may design appropriate masks to optimally sample the light field.
Figure 5-4: Optimal modulation of non-rectangularly bandlimited light fields (taken from [11]). (left) In this case the optimal mask RMF consists of unequally spaced impulses. (right) The light field convolves with the mask RMF to produce spectral replicas which line up along the horizontal slice without gaps. Note however that the spectral copies may overlap in other regions of the 2D Fourier space, since the sensor (red box) cannot detect those parts of the spectrum.

5.2.4 Obtaining Light Field Through Demodulation

The light field may be recovered through demodulation of the sensor slice (Figure 5-5). First, we take the Fourier transform of the raw sensor image, which is equivalent to the horizontal slice of the modulated light field spectrum. We then rearrange segments of the 1D sensor spectrum into their original $f_x, f_\theta$ locations to reassemble the 2D light field spectrum. Finally, the light field is obtained by taking the inverse Fourier transform of the 2D light field spectrum.

Figure 5-5: Demodulation involves reshaping the Fourier transform of the sensed image into the original 2D $f_x, f_\theta$ locations. The original light field may then be recovered by taking the inverse Fourier transform.
5.2.5 Extension to 4D Light Field Capture

Up to this point, the theoretical derivations were carried out in 2D for simplicity, but the foundations for 2D optical heterodyning are very much applicable in the general 4D case. By placing a 2D mask between the lens and the sensor, a bandlimited 4D light field may be modulated by convolving with the 4D mask RMF to produce spectral replicas in the 4D frequency domain. These spectral tiles encode the angular information in the 4D light field within the 2D spatial sensor space.

By extension of the optimal 1D mask pattern (Equation 5.8), the Fourier transform of the 2D mask would also contain a series of impulses along a tilted 2D plane, viz.

\[
C(f_1, f_2) = \sum_{i=-p_1}^{p_1} \sum_{j=-p_2}^{p_2} \delta(f_1 - i f_0^x)\delta(f_2 - j f_0^y)
\]  

(5.11)

where \( f_0^x \) and \( f_0^y \) denote the fundamental frequencies in the \( x \)- and \( y \)-dimensions of the mask. After modulation, such a 2D mask RMF produces \( t_1 \times t_2 \) spectral tiles, where \( t_1 = 2p_1 + 1 \) and \( t_2 = 2p_2 + 1 \). In the primal domain, the 2D mask is a sum of \( p_1 \) harmonic cosines in the \( x \)-direction and \( p_2 \) harmonic cosines in the \( y \)-direction.

As in the 2D case, recovering the 4D light field involves demodulating the sensor slice. We first take the Fourier transform of the 2D sensor image, then reassemble the 2D spectral tiles into a 4D array, and finally take the inverse Fourier transform (see Figure 5-9). Furthermore, as with all single-snapshot light field capture devices, there is a necessary tradeoff between spatial and angular resolution, since the sensor pixels are now used to sample the angular variation in rays. In other words, for a sensed image of \( N \times N \) pixels and \( t_1 \times t_2 \) angular resolution, the size of the resultant light field will be \( (N/t_1) \times (N/t_2) \times t_1 \times t_2 \).

5.2.6 Recovering Full-resolution Image

Despite the inherent loss of spatial resolution when acquiring light fields, optical heterodyning nonetheless allows the original full-resolution image to be recovered [10]. Naturally, the addition of a mask in the optical path casts soft shadows on the sensor and 'dapples' the captured photograph. In effect, these shadows represent the...
extent to which the scene irradiance is attenuated by the mask on average. Therefore, by compensating for the intensity variation within the bundle of rays reaching each pixel due to the mask, the ‘undappled’ image of the scene may be obtained (Figure 5-6).

Specifically, this is achieved by dividing the captured sensor image \( s(x, y) \) with a calibration image \( \gamma(x, y) \) of a uniform intensity Lambertian scene (e.g. a diffuse white plane), viz.

\[
I(x, y) = \frac{s(x, y)}{\gamma(x, y)}
\]  

(5.12)

Note that in-focus parts of the scene will remain focused in the compensated image \( I(x, y) \), while parts of the scene that were out of focus will remain blurred.

Besides recovering the full-resolution image, the same principle may also be used to correct for vignetting. In both the captured and calibration images, the effect of lens aperture-induced vignetting along the edges of the images is similar. Hence, by dividing the captured image by the calibration image, one may appropriately rescale the intensity at each pixel, so that the compensated image appears bright throughout.

Figure 5-6: Recovering full-resolution focused image of the scene by dividing the captured image by a calibration image (taken from [10]). The inset shows the shadows cast by the mask (top) and the corresponding part of the compensated image (bottom).

5.3 Practical Implementation

From the above analysis, the optimal mask for optical heterodyning is a 2D cosine mask of multiple harmonics, placed at the correct distance between the lens and the
Figure 5-7: Zoomed-in view of sum-of-cosine mask (taken from [10]). The plot on the right visualizes a slice along the 2D mask as a sum of 4 harmonic cosines ($f_0 = 1$ cycle/mm) and a DC term. Note that the resultant sum is always non-negative.

sensor. However, there are several practical issues in building such a heterodyne light field camera. We first highlight these issues before describing two prototypes that Veeraraghavan et al. have constructed in [10] and [11].

5.3.1 Practical Considerations for Mask

Creating a sum-of-cosines mask is done by printing a pattern on a transparency. Since the printed mask does not allow for negative values, the DC component of the sum-of-cosines must be boosted to keep the pattern in the positive range. Figure 5-7 shows a portion of a cosine mask used in [10] with a 1 cycle/mm fundamental frequency and four harmonics in both dimensions. This mask allows for a 9-by-9 angular resolution in the light field.

The process of printing is also relatively cheap yet high-quality. Each mask may be printed for just a few dollars. Moreover, current printing technology allows the mask pixel size to reach as low as 25nm with 1024 grayscale levels [11]. This resolution is sufficient for most standard-size CCD sensors, although presumably for smaller sensors the attenuating elements might become so fine that diffraction effects start to set in.

A final issue pertains to the placement of the mask within the camera. Although the position and tilt of the mask may be precisely varied, the calculated optimal mask position might not always be practically realizable, especially for smaller cam-
eras. With a smaller lens aperture size and shorter lens-sensor distance, the optimal $\alpha$ is sometimes so small that the mask has to be placed right next to the sensor. Unfortunately, most CCD sensors have a sheet of glass over them to protect the internal circuitry, thus preventing the mask from being optimally positioned. To circumvent this problem, custom sensors may have to be manufactured with the mask-printing incorporated into the fabrication process.

5.3.2 Design of Heterodyne Camera

With the above issues in mind, Veeraraghavan et al. successfully constructed two prototypes of the heterodyne light field camera. The first prototype in [10] was built using a flatbed scanner with a letter-sized patterned transparency inserted behind the scanner glass surface (Figure 5-8(a)). The scanner served as the imaging sensor of a large-format camera, fitted with a 210mm $f/5.6$ Nikkor-W lens. However, the jagged motion of the scanner sensor introduced significant noise into the acquired images and the reconstructed light fields. Moreover, due to the large-format set-up, there was significant vignetting during image capture, further reducing the quality of the light field datasets.

To alleviate these problems, a more elegant version was constructed using a Mamiya 645ZD medium format digital camera fitted with a 210mm Mamiya Sekor lens, as shown in Figure 5-8(b) [11]. Such a camera not only has the advantage of being hand-held, but also has an easily accessible sensor, compared to conventional cameras in which the sensors are obstructed by optics and electronics. As such, the camera can be easily modified by placing the mask over the glass sheet covering the sensor, giving a mask-sensor distance of 1.2mm. Different cosine masks were printed with 3, 4 and 5 harmonic frequencies at high resolution using a Kodak Light Valve Technology (LVT) continuous tone film recorder. In particular, the mask with 3 harmonics produced $7 \times 7$ angular samples with a spatial resolution of $237 \times 175$.

Camera images are captured as raw MEF files to avoid compression and subsequently converted to 16-bit linear images for processing. Post-capture image processing follows the same procedure outlined in Section 5.2.5, with the addition of
Figure 5-8: (a) First prototype of the heterodyne light field camera, as described in [10]. A 2D cosine mask is placed behind the glass surface of a flatbed scanner. (b) Hand-held Mamiya medium format digital camera retrofitted with a cosine mask [11]. The mask is placed on top of the sensor and held in place using the IR filter from the camera. After this simple modification, the back of the camera is reattached to the body.

5.4 Advantages and Limitations

Light field acquisition via optical heterodyning has several advantages. First, the addition of a mask may be easily carried out since they are non-refractive and do not require the same level of accuracy as other microlens-based designs (e.g. in [4]). The low printing cost of masks and ease of retrofitting an existing digital camera
Figure 5-9: Summary of 4D light field acquisition using a heterodyne camera. The mask used to modulate the image consists of 3 harmonic frequencies of 8, 16, 24 cycles/mm, giving $7 \times 7$ angular samples with a spatial resolution of $237 \times 175$.

Figure 5-10: (a) Effect of vignetting compensation on light field images. Both images depict the same subset of the light field, taken from the bottom left corner of the lens aperture plane. After dividing each light field view by the corresponding calibration light field image, the brightness and contrast of the images drastically improves. (b) Two of the $11 \times 11$ light field views from this dataset, taken from diametrically opposite ends of the lens aperture plane. The dotted line clearly shows the parallax between these two views, as the tip of the letter 'A' in the bottom image is displaced to the right relative to the top image.
also suggest that the addition of a mask will not significantly increase manufacturing cost and complexity. Users may even replace the mask in the camera depending on their needs. Finally, mask-based cameras retain the ability to obtain full-resolution images from the captured image, which other plenoptic camera designs are unable to accomplish.

However, there are also numerous limitations to mask-based optical heterodyning. The main disadvantage of using masks is the loss of light (by ~1 f-stop) due to their attenuating effect. Consequently, SNR is lowered and exposure time must be increased to make up for this, potentially leading to motion blur.

Second, when the bandlimited light field assumption fails, the reconstructed light fields suffer from aliasing effects [12]. If the angular \( f_\theta \) bandlimit is not true, then a sum-of-cosines mask designed for rectangular bandlimits can only modulate the light field up to the angular bandwidth \( f_{\theta 0} \), and frequencies above \( f_{\theta 0} \) will not be recovered. This leads to smoothing along the angular \( \theta \) dimension. On the other hand, if the spatial \( f_x \) bandlimit is false, then energy in the higher spatial frequencies of the light field overlap with the lower angular dimensions. This prevents the accurate reconstruction of light field. To prevent aliasing, the light field must be treated with suitable pre-filters so as to bandlimit it.

Finally, since masks attenuate the light entering the camera, they effectively reduce the lens aperture size and increase the blurring effect of diffraction. As such, masks might not be suitable for diffraction-limited optical systems such as microscopes. Moreover, as the attenuating elements of the mask become smaller, more diffraction artifacts are introduced, potentially limiting the maximum resolution at which light fields may be reconstructed even with higher sensor resolution.
Chapter 6

Depth Reconstruction

Having examined various methods of acquiring light fields from a single snapshot as well as multiple images, we proceed to utilize these information-rich arrays to reconstruct the 3D depth of a scene and the shape of objects within. This has important applications in 3D surface imaging and machine vision. In this chapter, we describe two algorithms for recovering depth information, namely depth-from-synthetic aperture focus and depth-from-stereo. We also present an experimental comparison of these two methods on light fields obtained from a camera array and heterodyne camera.

6.1 Overview of Depth Estimation Techniques

Getting 3D depth from images is a classic problem in computer vision, and a plethora of techniques already exist for this purpose. Most work in this field has focused on stereo vision involving two images. Even from a single image, it is also possible to infer depth through a probabilistic learning approach [14].

For a collection of light field images, perhaps the most straightforward technique is to match objects across different views using cost functions based on Sum of Absolute Differences (SAD), Sum of Squared Differences (SSD) and cross-correlation, among others. Indeed, such a block-matching approach has been used by Adelson and Wang [1] as well as Frigerio [18] to estimate depth from plenoptic cameras and
Figure 6-1: Point-plane correspondence: Every point in the scene is represented by a line in the 2D $s-u$ slice of the light field and a plane in the general 4D space. The orientation of the plane depends only on the depth of the point.

Another way to find depth from light fields is by exploiting the point-plane correspondence property. From Figure 6-1, we see that for any point on the $s-t$ plane, rays coming from a light source will intersect the $u-v$ plane at only one point. This implies that a scene point corresponds to a line in the 2D $s-u$ slice, or a plane in the 4D light field space. Since the slope of this plane is related to the depth of the scene point, gradient-based methods may be employed towards range-finding [19].

The two approaches adopted in this chapter hinge on the notion that the multiple viewpoints from a light field may be meaningfully combined to give an aggregate measure of scene depth. The intuition behind this is that even if an object is occluded in some views, other viewpoints that see the object may still be used to reconstruct its depth. Such techniques thus have an edge over block-matching algorithms, which mostly require the object to be visible in all views.

The first method involves projecting the light field images onto a synthetic focal plane and taking their average. In the aggregated image, points lying on this focal plane will appear sharp, while all other points will be blurred, much like an image from an ordinary lens. This technique is called *synthetic aperture focusing* [6], because each light field view represents a point on a synthetic lens, whose focal length may be altered by warping the views. We may then apply a sharpness criteria to search for the depth at which an object appears most in-focus.
The second approach resembles multi-baseline stereo in that the disparity which produces the greatest degree of color constancy indicates the object depth. Since the light field images are considered as a whole rather than individually compared to a reference image, the effect of outliers due to occlusion is reduced.

The specific algorithms used for depth-from-focus (DFF) and depth-from-stereo (DFS) are described later on in Section 6.3.2.

6.2 Synthetic Aperture Focusing

As both depth-from-focus and depth-from-stereo methods rely on projective warps of synthetic aperture views, we shall now formally introduce the concept of Synthetic Aperture Focusing.

Consider the case of a dense array of pinhole cameras taking pictures of a scene from different viewpoints. In the general configuration, each image is originally taken on a different film plane, and must thus be projectively warped onto a common plane for any meaningful comparison (Figure 6-2(a)). The result is that pixels corresponding to objects located on this plane will reinforce each other, while those off the plane will be blurred out due to parallax. Hence, the synthetic aperture is effectively focused on this reference plane. Formally speaking, such a projective warp is known as a homography. Creating a synthetically focused image for a different focal plane then requires applying the appropriate homographies to the camera images and subsequently computing the average [7].

However, the process of refocusing is greatly simplified if the cameras already lie in a plane and their images have been projected onto a parallel plane via an initial homography (Figure 6-2(b)). In this case, objects not on this reference plane are still displaced from each other due to parallax, but we may focus on any frontoparallel plane by simply shifting the images by the amount of parallax and then adding them together [8].

43
Figure 6-2: (a) In the general configuration, synthetic aperture focusing is achieved by projectively warping camera images onto a reference plane \( \Pi_0 \) and averaging them. Point \( Q \) on the plane is in-focus, while Point \( P \) lying on another plane \( \Pi \) is blurred due to parallax. To focus on \( \Pi \) different homographies must be applied to the camera images [7]. (b) For the case of a planar camera array with images that have already been projected onto a parallel reference plane \( \Pi_0 \), synthetically focusing onto different frontoparallel focal planes is then a simple process of shifting and adding images to negate their parallax. [7]

### 6.2.1 Theoretical Analysis of Refocusing

Analytically, the synthetic aperture may be thought of as a lens \((u-v)\) forming an image on a synthetic film plane \((x-y)\) using the two-plane parametrization for light fields. The image that forms inside this synthetic aperture camera is the integral projection of the light field radiance arriving at every pixel, viz.

\[
E_F(x, y) = \frac{1}{F^2} \int \int L_F(x, y, u, v)A(u, v) \cos^4 \phi dudv, \tag{6.1}
\]

where \(E_F(x, y)\) is the image intensity at \((x, y)\) for a lens-film distance of \(F\), \(A(u, v)\) represents an aperture modulation function, and \(\phi\) is the angle between the ray \(L(x, y, u, v)\) and the film plane normal [5]. If we use the paraxial approximation to eliminate the \(\cos^4 \phi\) term, neglect the constant \(\frac{1}{F^2}\) and also absorb \(A(u, v)\) within
Figure 6-3: Moving the synthetic film plane from \( F \) to \( F' \) only involves reparametrizing the light field [5].

Let \( L_F \) itself, Equation 6.1 may be simplified to

\[
E_F(x, y) = \int \int L_F(x, y, u, v) dudv \tag{6.2}
\]

Now, from Figure 6-3, we see that moving the film plane from \( F \) to \( F' \) leads to a simple reparametrization of coordinates. By identifying similar triangles, the ray that comes from the point \( u \) on the synthetic lens aperture and reaches the new focal plane \( F' \) at \( x \) also intersects the original focal plane \( F \) at \( u + \frac{x-u}{\alpha} \), where \( \alpha = F'/F \). The radiance reaching the new focal plane is then related to the original light field by the following expression: [5]

\[
L_{F'}(x, y, u, v) = L_{(\alpha F)}(x, y, u, v) = L_F(u + \frac{x-u}{\alpha}, v + \frac{y-v}{\alpha}, u, v) \tag{6.3}
\]

Combining Equations 6.2 and 6.3 yields the synthetically refocused image at the new focal plane \( F' = \alpha F \):

\[
E_{(\alpha F)}(x, y) = \int \int L_F(u + \frac{x-u}{\alpha}, v + \frac{y-v}{\alpha}, u, v) dudv \tag{6.4}
\]

As such, Equation 6.4 shows that synthetic refocusing in the case of two parallel planes is equivalent to summing shifted versions of all the \((u,v)\) subaperture views (i.e. the individual projected camera images), as intuitively predicted.
6.2.2 Calibrating Parallax for Planar Camera Arrays

Given that the images from a planar camera array may be shifted and summed to vary the focus, implementing synthetic aperture focusing thus requires precise quantification of the parallax at each focal depth. As we shall see, this involves computing the relative camera positions and using the information to calibrate the parallax for any desired depth.

To examine the relationship between parallax, depth and camera geometry, we consider a planar array of cameras $C_0, \ldots, C_m$ initially focused on a parallel reference plane, as depicted in Figure 6-4. This may be done by aligning features on a calibration grid parallel to the camera plane. Now, any point $Q$ on the reference plane has the same pixel coordinate in all images, and thus appears focused in the aggregated image. However, a point $P$ not on the reference plane is projected to two different points in two images. From similar triangles, this parallax $\Delta p_i$ is given by

\[
\Delta p_i = \Delta x_i \frac{D_p - D_0}{D_p},
\]

(6.5)

where $D_0$ and $D_p$ denote the distance from the camera plane to the reference plane and an arbitrary focal plane $\Pi$ respectively. In other words, the parallax of a given camera $C_i$ with respect to a reference camera $C_0$ is the product of its relative camera displacement $\Delta x_i$ and relative depth $d_p$, defined as $\frac{D_p - D_0}{D_p}$ [8].

Equation 6.5 may be applied to multiple camera views of the same point $P$ to produce a vector of parallax measurements, viz.

\[
\begin{bmatrix}
\Delta p_1 \\
\vdots \\
\Delta p_m
\end{bmatrix} = \begin{bmatrix}
\Delta x_1 \\
\vdots \\
\Delta x_m
\end{bmatrix} \cdot d_p.
\]

(6.6)

Hence, if the exact depth of point $P$ is known, the vector of camera displacements with respect to the reference camera may be determined from the parallax vector via a simple scalar operation.

Recovering the camera positions in this manner is important for two reasons.
Most obviously, once we have the camera positions in hand, we may then compute the parallax of each camera for any desired depth, without having to perform a new calibration each time. This allows us to compensate for the parallax accordingly and carry out synthetic aperture focusing. A second and more subtle reason is that the actual camera geometry is often unknown and may be difficult to measure accurately. Even if this could be done, the camera positions are found on a different scale from that of the reference plane. Hence, an additional coordinate-mapping transformation must be applied before computing image shifts. In contrast, parallax measurement of a point off the reference plane robustly gives us the relative camera positions in terms of the coordinate system on the reference plane, thus permitting the direct use of Equation 6.6.

The above observations suggest a simple procedure for synthetic aperture focusing using a planar array. First, we focus the cameras onto a frontoparallel plane using homographies derived by imaging a calibration grid. Then, we compute the relative camera positions from parallax measurements of one or more scene points. Finally, the camera positions are used to vary the focal plane over a range of depths. This
amounts to shifting the image from camera $C_i$ by $-\Delta p_i = -\Delta x_i d_p$ and averaging the shifted images.

6.3 Experimental Methodology

Based on the above method for camera calibration and synthetic aperture focusing, we developed depth-from-focus and depth-from-stereo algorithms and tested them on real light fields to evaluate their performance.

6.3.1 Data Acquisition

Light field datasets were obtained from a planar camera array and a mask-based heterodyne camera. The camera array used was the ProFUSION 25 camera developed by ViewPLUS Inc. As briefly mentioned in Section 4.1, the ProFUSION 25 is an array of $5 \times 5$ VGA cameras with a 12mm spacing between cameras (Figure 6-5(a)). This allows the camera array to capture a scene from 25 different viewpoints. Each individual camera has a 5.34mm f/2.8 microlens with a field of view of 36.8°. Images were stored as 8-bit color or grayscale PPM files at a resolution of $640 \times 480$. ViewPLUS Inc. has also provided software to capture raw camera images as well as perform real-time synthetic aperture refocusing on a PC, which we have used to directly acquire light fields.

To calibrate the camera array and recover the relative camera positions, we used two letter-sized grids consisting of alternating black-and-white rectangles and a barcode pattern mounted on a flat, black panel. Image correspondences were obtained by applying an efficient subpixel image registration routine that matches features by cross-correlation (see Figure 6-5(b)) [15]. The central camera (camera 12) was chosen as the reference view.

First, the calibration target was placed parallel to the camera at a distance of 80cm. This plane was defined as the reference plane $\Pi_0$. Using the image registration algorithm, each camera image was shifted to align onto this focal plane. Next, the calibration target was moved onto another parallel plane $\Pi$ at 140cm from the camera.
Figure 6-5: (a) The ProFUSION 25 camera array was used to acquire light fields for demonstrating synthetic refocusing and depth acquisition. The central camera (camera 12) is chosen as the reference view. (b) A shifted calibration image stacked on top of the reference view, after applying an image registration algorithm. The algorithm robustly matches corners and edges and outputs the subpixel shift required. Points lying on the reference plane overlap perfectly and are in-focus, while ghosts arising from parallax may be seen for objects not on the focal plane.

By applying the image registration algorithm once more, the amount of shift for each raw camera image was again computed. The parallax of objects on $\Pi$ relative to the reference plane $\Pi_0$ may then be measured by subtracting the initial shift required to focus on $\Pi_0$ from the new shift for $\Pi$. Thereafter, the relative camera positions was recovered using Equation 6.6, so that synthetic aperture focusing may be performed. Interestingly, despite knowing the rough geometry of the ProFUSION camera, this calibration step was found to be crucial, as the raw captured images were irregularly spaced apart.

The ProFUSION camera was then used to capture light fields of numerous scenes in a straightforward point-and-shoot manner. In the rest of the chapter, however, we will only present a few light fields which capture the salient performance characteristics of our algorithms.

Additionally, the mask-based heterodyne camera, which was described in Section 5.3.2, was used to capture light fields for evaluation. The goal of using the heterodyne camera was twofold: First, we carry out synthetic aperture refocusing for this camera
to demonstrate the utility of heterodyning. Second, we wish to implement our depth-from-stereo and depth-from-focus algorithms in the context of a real-aperture camera, as compared to a camera array behaving as a synthetic lens. Hypothetically, the smaller aperture size of a real lens could place limitations on the applicability of our algorithms.

The heterodyne camera was fitted with a cosine mask of 5 harmonic frequencies to capture a single photograph of a scene, as well as a calibration image for vignetting correction. Through the technique described in Chapter 6, this image was decomposed to produce a light field with a directional resolution of $11 \times 11$ and a spatial resolution of $191 \times 257$, on which our algorithms were directly applied. Unfortunately, no calibration was made against the physical geometry of a scene, and thus quantitative depth information cannot be extracted. Nonetheless, qualitative performance may still be inferred.

### 6.3.2 Description of Focus and Stereo Algorithms

We now describe the computational steps involved in shifting raw images, and obtaining depth-from-focus and depth-from-stereo.

In synthetic aperture focusing, shifting the raw images is a special and simple case of a homography, which may be applied by matrix-multiplying a projective transform. However, we achieve the same effect with greater computational speed by padding the image to be shifted with the edge pixel values and then cropping down to size. This allows for pixelwise displacements, and so the required shift computed from Equation 6.6 is rounded to the nearest integer pixel.

In Depth-from-Focus, a sequence of images focused at a range of depths is computed through synthetic aperture focusing. In these refocused images, scene points that are out-of-focus are blurred while those that are in-focus appear sharp. The degree of sharpness of a pixel may be measured by the variance of the neighborhood around it, where a higher variance denotes greater sharpness. Specifically, we scan a patch of size $5 \times 5$ pixels around the stack of refocused images. Each pixel is then assigned the depth corresponding to the refocused image in which the patch variance
is maximum. In this way, a depth map (which contains the depth value for each pixel) may be computed. Furthermore, since the patch variance also gives a confidence estimate of the sharpness, pixel depth values with a maximum variance below a certain threshold are ignored. Low confidence depth values typically correspond to regions that are smooth and featureless, while confidence is high around edges and other regions of great spatial change. Neglecting low confidence points improves the quality of the depth map.

In Depth-from-Stereo, a range of depths are again considered. For each candidate depth, each light field view is shifted by its respective disparity (as given by Equation 6.6) and the variance of rays from this stack of shifted images is calculated. If a scene point is in-focus at this particular depth, the pixel value for all the shifted images will agree substantially with one another since all rays have effectively converged onto the same point, leading to a low variance. Conversely, if the scene point is out-of-focus at this depth, then the same pixel coordinate in different shifted images would correspond to different scene objects, thus resulting in a high variance. As such, the depth of each pixel is taken to be that at which its variance in the shifted image collection is minimized.

In the 1D case, the objective function may be mathematically expressed as

\[
v_d(x) = \frac{1}{N} \sum_i (I_{i,d}(x) - \bar{I}_d(x))^2, \quad (6.7)
\]

where \( I_{i,d}(x) \) denotes the image from camera \( i \) projected onto the plane at depth \( d \), and \( \bar{I}_d(x) \) represents the mean of the warped images from the \( N \) cameras (also equivalent to the synthetically focused image at \( d \)). By minimizing the objective function \( v_d(x, y) \) for a \((x, y, d)\) triplet, a depth map may be constructed from stereo.

In both algorithms, the depth range and depth resolution was user-defined based on the scene. A median filter of \( 5 \times 5 \) window size was also applied whenever appropriate, so as to reduce speckle noise yet preserving edges. All computation was done in MATLAB.
6.4 Results

6.4.1 Digital Refocusing

Figure 6-7 shows the results of synthetic aperture refocusing for a complex light field (subsequently referred to as 'books') taken using the ProFUSION 25 camera. The scene contains many objects, which increases the chance of occlusion across different camera views. However, as the focal depth is increased away from the camera plane, different objects in the scene come into focus. In effect, the planar camera array simulates the defocus of an ordinary camera lens with an aperture size equal to the size of the array. Due to the large effective aperture size, the depth of view for each refocused image is shallow, enabling only objects at a particular focal depth to appear sharp while blurring all other objects. Nonetheless, the full effective depth of field remains wide, since the camera array may synthetically refocus on objects in both the foreground and the background.

Furthermore, Figure 6-7 demonstrates the ability of synthetic apertures to reconstruct occluded surfaces. In the reference view (Figure 6-6), the title of the Linear Algebra book in the background is occluded by a pair of batteries. However, when the camera array is synthetically focused onto the plane of the book, the words come into clear view. Such reconstruction is made possible by the large baseline of the synthetic aperture, such that we can capture enough rays that go around the foreground occluders to image the objects behind. This advantage of synthetic aperture focusing clearly has potential applications in surveillance and military settings.

Synthetic aperture refocusing by integral projection is also possible for the heterodyne camera, as shown in Figure 6-9. This is done after carrying out a vignetting correction to compensate for the light loss around the fringes of a refocused image, due to the simulated aperture (Figure 6.4.1). By shearing and warping the light field views obtained from the camera, we may bring either the toy in the foreground or the book in the background into sharp focus. Given that the original captured image was only focused on the toy, synthetic aperture refocusing may be used to extend the depth-of-field of cameras through post-capture processing.
Reference view of the ‘books’ light field. This light field contains many non-ideal features, such as occlusions, specular reflections and insufficient texture, and is thus a good test-bed for our algorithms.

Figure 6-7: (clockwise from top right) Refocused photographs at increasing depth away from the ProFUSION camera. Different objects in the scene at come into focus at different depths. Notice that the title of the Linear Algebra book in the inset may be clearly seen in the refocused image, even though it was occluded in the reference view.
Figure 6-8: (a) Refocused image from heterodyne camera without vignetting correction. Notice the dark fringe around the sides, simulating the vignetting effect of a lens (b) Same refocused image with vignetting correction. The brightness and contrast improves.

Figure 6-9: (a) Low resolution image refocused onto the toy in the foreground (b) Low resolution image refocused onto the book in the background.

With a directional resolution of $11 \times 11$, synthetic refocusing would ideally match an aperture 11 times narrower [4]. As an extension, given the f/number of the heterodyne camera lens, the extent of blur in the refocused images may be compared against an actual camera with a f/number 11 times larger using resolution charts, so as to quantify the losses due to diffraction and attenuation from the mask.

### 6.4.2 Depth Estimation

The depth maps for the ‘books’ light field dataset computed using our depth-from-focus and depth-from-stereo algorithms are shown in Figures 6-10 and 6-11 respectively. The depth resolution was chosen to be 2cm for near objects and 5cm for far objects. A comparison with the manually derived ground truth showed that both
methods could accurately estimate the depth of objects to within this resolution. As such, DFF and DFS algorithms serve as viable range-finding tools.

However, a closer look at the depth-from-focus depth map (Figure 6-10) reveals several limitations with the method. First, only edges are represented clearly with high confidence; most parts of the scene lack sufficient texture for DFF to robustly reconstruct their depths and are thus ignored, resulting in the large patches of dark blue. This may be explained by recalling that DFF seeks to maximize the variance of the neighborhood around a pixel. For regions with insufficient texture, the pixel values for differently refocused images are likely to be similar and thus undetected by DFF.

Furthermore, DFF fails in the case of specular reflections and other non-Lambertian (shiny) surfaces (Region (2)). Each individual camera in the array perceives reflections differently due to their varying viewpoints, and thus the camera cannot be tuned to converge rays onto the source of reflection. As seen from the bottom right inset, the variance for a wide range of depths are close to each other, making it difficult to discern the true depth from DFF.

A last point of interest pertains to the occlusion of the Linear Algebra book by the pair of batteries in Region (3). Intuitively, we know that the correct depth at Region (3) should be that of the batteries, yet DFF returns the false depth of the lettering on the book. As noted by Schechner et al. in [9], such a phenomenon occurs when the point is severely occluded and the chief ray from the object does not reach the lens, as in the case of the book. Hence, during the focus search, the same point indicates a focused state when both the book and the batteries are in-focus. Thus, simple DFF returns a double-valued depth at Region (3).

Turning to the depth-from-stereo depth map (Figure 6-10), we note improvements in depth estimation for DFS over DFF, especially for textured regions bounded by edges, such as the planar surfaces of the books (in turquoise and yellow). In particular, DFS produces an accurate depth for a wider area of the barcode pattern in Region (1) of both depth maps. While both DFS and DFF may not compute the correct depth (as shown in the top right inset of Figures 6-10 and 6-11), DFF meets with
substantially more noise in its patch variance at lower candidate depth levels. At first glance, the failure of DFF at Region (1) might seem peculiar, given that the barcode pattern is a well-defined texture that should be easy to distinguish by the focus setting. However, as Vaish et al. prove in [6], certain textured surfaces that lack sufficiently high-order spatial derivatives cannot be reconstructed accurately by focus because the mean image does not vary sufficiently with depth. The barcode pattern is likely to belong to this category.

Like DFF, DFS also runs into problems with insufficient texture and specular reflection, as evidenced by the top right hand corner of the black calibration panel and Region (2) being reconstructed incorrectly (see bottom right inset). This occurs for the same reason as described earlier.

Finally, the edges of objects in the DFS depth map are not well-defined but instead appear to blend in with the neighboring objects, in contrast to the accurate edge recovery using DFF. The heat gun is a good example of this effect. The DFS perceived depth value falls off gradually along edges because at these regions, the sides of the foreground object as well as the object behind are occluded in some views but not in others. The variance will thus not reach a minimum at the correct depth unless the occluder is of a fairly uniform appearance [6].

Next, we consider the depth maps computed for the light field dataset taken by the heterodyne camera. Due to the lack of calibration, the depth maps encode depth in terms of disparity instead of physical distance, and were quantized to 20 depth levels. As shown in Figure 6-12, both DFF and DFS produce similar-looking depth maps. Since there is less occlusion in this scene, both DFF and DFS are relatively robust along edges. Also, as before, both DFF and DFS have difficulty with untextured surfaces, although DFS is better able to make out surfaces. From the DFS depth map, we can tell that the toy elephant’s trunk and foot is closer to the camera than the rest of its body.

The ability to compute depth even for a small baseline lens again points to the utility of synthetic aperture-based algorithms. One way to make use of the depth maps obtained is to create an all-in-focus image whereby the intensity of each pixel
Figure 6-10: Depth map from depth-from-focus algorithm without median filtering. Dark blue regions indicate areas that were ignored due to the confidence threshold. Under DFF, the edges of objects are mostly correctly reconstructed, but other regions which lack sufficient texture are not. DFF also fails when there are with specular reflections (2) or no high-order spatial derivatives (1).

Figure 6-11: Depth map from depth-from-stereo algorithm with median filtering. Compared to DFF, the depths of surfaces are better constructed. However, DFS also has difficulties with specular reflections and occlusions.
is taken from the refocussed image corresponding to its depth. However, the smaller baseline also places limits on the accuracy of DFF and DFS methods, which we will attempt to quantify subsequently.

6.4.3 3D Surface Imaging

Finally, since depth-from-stereo outperforms depth-from-focus for textured regions, we make use of the algorithm to reconstruct the surface of a patterned mug, as shown in Figure 6-13. Using a depth resolution of 0.5cm, the 3D surface plot shows the cylindrical geometry of the mug and also yields a mug diameter that roughly agrees with ground truth. Nonetheless, strict thresholds and a median-filter of window size $7 \times 7$ had to be applied to the depth map before such a 3D geometry was obtained. Presumably, depth accuracy and the smoothness of the surface may be increased by upsampling the light field and acquiring more camera views.
Figure 6-13: (a) Reference view of the light field of a mug. (b) 3D surface visualization of mug for the region bounded by the orange box.
6.5 Discussion

6.5.1 Comparison of Depth-from-Focus and Depth-from-Stereo

In summary, our results indicate that depth-from-focus is more robust for edges, while depth-from-stereo is more robust for most unoccluded textured surfaces. Both methods fail under non-ideal conditions of specularities, occlusion and textureless surfaces. However, the two-dimensionality of the synthetic aperture permits robust depth recovery by allowing more image points to contribute to the depth estimation and a higher SNR ratio resulting from the increased amount of light entering the aperture.

In comparing between the two methods, Vaish et al. also observe that depth-from-focus is less sensitive to sensor noise and camera bias due to the averaging effect of refocusing [6]. Furthermore, Vaish et al. show that for sufficiently textured surfaces, DFF performs better than DFS as occlusion increases [6].

6.5.2 Performance Potential of Algorithms

It is also worth noting that our algorithms are ultimately limited by the sampling of rays afforded by the ProFUSION camera and heterodyne camera. If the light field is not sampled densely enough, we observe focus-like effects even in supposedly defocused regions, thus leading to inaccuracies in depth estimation. Moreover, the relatively small effective aperture size of the ProFUSION camera not only implies a wider refocusing depth-of-field and thus lower depth resolution for DFF, but may also limit the sensitivity of disparity measurements in DFS.

To further understand the algorithms developed, we shall now briefly examine their performance potential. At the heart of both DFF and DFS is the image shift that corresponds directly to the object depth. This is given by Equation 6.5, reproduced here for convenience:

$$\Delta p_i = \Delta x_i \frac{D_p - D_0}{D_p}, \quad (6.8)$$

The system sensitivity to depth may be found by taking the derivative of Equation
6.8 with respect to object depth, viz.

\[
\frac{d(\Delta p_i)}{d(D_p)} = \frac{D_i}{D_p^2} \Delta x_i
\]  

(6.9)

Based on Equation 6.9, we see that a larger camera displacement or equivalently, a larger baseline, leads to greater sensitivity to depth. Also, as the target depth increases, sensitivity falls off and approaches zero. This implies that for increasing depths, a larger change in target depth will gradually require smaller disparities. Conversely, the smallest change in disparity (limited by the pixel-wise shifting code) places a limit on depth resolution, depending on the target position. As such, to improve the performance of DFF and DFS algorithms, the baseline of the light field camera should be enlarged, and accurate subpixel image shifts must also be included. The approach taken in this chapter did not fulfill either condition, and thus can be further modified.

As a further extension, we can consider projecting the images onto focal surfaces of arbitrary shape instead of considering only frontoparallel planes. This requires computing more complex homographies for refocusing, but would also intuitively lead to more accurate depth reconstruction if the arbitrary focal planes matched well with the actual scene geometry. To mitigate the effects of occlusion, we may also adopt a layered scene model and modify our objective functions to give more weightage to unoccluded rays.
Chapter 7

Conclusions

This thesis has described the theory, acquisition and practical application of light fields. In this final chapter, we will summarize the key takeaways and propose directions for future work.

As described in Chapter 2, the full complexity of the structure of light permeating a scene may be condensed into a 4D light field array. In essence, the light field contains the radiance along all rays in free space. It is typically parameterized in the two-plane parametrization for computational convenience.

Chapters 3-5 then described several methods to acquire light fields. Active Wavefront Sampling, which was described in Chapter 3, uses an eccentric rotating aperture to sample the wavefront impinging on the main lens, track the motion of image points and thus determine the depth of scene features. By taking a sequence of images for different subaperture openings over time, a light field is effectively obtained. Such a technique has found great application in 3D surface imaging, due to its use of only a single lens and ability to obtain depth in real-time.

Nonetheless, it might be preferable to acquire all the information from the scene with a single snapshot. Chapter 4 described two methods of doing so. A straightforward way to accomplish this would be to use a dense camera array. More elegantly, one may also use a hand-held plenoptic camera, which uses a microlens array to rebin the light rays reaching the imaging sensor in order to retain directional information. The structure of light is then encoded within a single photograph.
Chapter 5 described a non-refractive heterodyne light field camera, which avoids the microlens-related problems of aberrations and misalignment. By inserting a patterned cosine mask between the lens and the sensor, the angular dimensions of the light field that are normally cut off from the sensor are modulated back onto the sensor space. Furthermore, such a camera retains the ability to recover full-resolution images. This mitigates the inherent tradeoff between spatial and angular resolution faced by all single-snapshot light field capture devices. However, due to the attenuating nature of the mask, SNR is lowered and diffraction effects are amplified.

Finally, using light fields obtained from a commercial planar camera array and the heterodyne camera, we developed algorithms to perform synthetic aperture focusing and extract depth from focus and stereo measures. It was shown that the light field images from both cameras may be shifted appropriately and averaged so as to refocus onto a new focal plane. This allowed views around occluded objects to be reconstructed. Furthermore, depth-from-stereo and depth-from-focus methods both yielded accurate depth estimates within the user-defined depth resolution. Another advantage of these methods is that they are more robust than typical block-matching methods in the presence of occlusions, since they aggregate the information from multiple light field views instead of comparing them to each other. Comparing between the two algorithms, depth-from-focus is more robust along edges, while depth-from-stereo performs better for sufficiently textured surfaces. A brief quantization of the performance limits of both methods was described, based on the relation between disparity and target depth.

7.1 Future Directions

As shown in this thesis, several different methods are available for light field acquisition. Future work could begin by formulating a unified framework for examining these methods. For instance, the eccentric aperture of the AWS set-up and the cosine mask in the heterodyne camera are both examples of non-refractive ray modulators. Presumably, there exists some correlation between the two mechanisms for obtaining
light fields, particularly in the frequency domain.

Another synergistic option between AWS and optical heterodyning takes the form of developing 4D universal ray modulators that are angle and location sensitive. This may be practically implemented by using a hologram mask or building a combination of a color LCD filter without the backlight and a lens relay system. With 4 degrees of freedom in both angular and spatial dimensions, such a ray modulator may allow for more information to be captured. A further advantage of using an LCD filter is that the mask pattern may be readily changed, allowing different resolution light fields to be captured. A study should be made to determine the optimal sampling pattern for such a versatile ray filter.

A third possibility exploits the time dimension. On top of heterodyning in angle and space, one might also encode the exposure time of images to attain a fifth degree of freedom. Moreover, by taking a sequence of mask-modulated images, the low-resolution light fields obtained may be combined to yield greater resolution.

In our work on synthetic aperture focusing, we have shown that the exact geometry of light field camera positions need not be known accurately prior to processing. By applying a plane-parallax calibration, depth information can still be extracted from the light field. This is in contrast to the AWS methodology, in which the position of the off-axis sampling point may be precisely known so as to perform accurate block-matching. It remains to be seen if depth-from-stereo and depth-from-focus methods, which aggregate the light field views instead of matching them, can also be applied to AWS datasets with good effect. At the same time, the performance limits to synthetic aperture-based depth estimation methods should be further probed. For instance, if the subaperture views were known accurately, as in the case of AWS, then DFF and DFS methods might yield better accuracy and resolution.

Finally, because light fields remain in the realm of ray optics, wave optics effects such as diffraction and aberrations are not easily taken into account. Nonlinearities like specularities also distort the sampling of rays from different light field views. Perhaps an augmented theoretical framework for light fields can be incorporated into future studies of depth reconstruction.
Bibliography


