Designing miniature computational cameras for photography, microscopy, and artificial intelligence
by
Vivek Boominathan
A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree
Doctor of Philosophy

APPROVED, Thesis Committee:

[Signatures]

Ashok Veeraraghavan, Chair
Assistant Professor of Electrical and
Computer Engineering, Rice University

[Signatures]

Jacob Robinson
Assistant Professor of Electrical and
Computer Engineering, and
Bioengineering, Rice University

[Signatures]

Richard Pariniuk
Victor E. Cameron Professor of Electrical
and Computer Engineering, Rice
University

[Signatures]

Anastassia Kyrillidis
Assistant Professor of Computer Science,
Rice University

Houston, Texas
May, 2019
ABSTRACT

Designing miniature computational cameras for photography, microscopy, and artificial intelligence

by

Vivek Boominathan

The fields of robotics, Internet of things, health-monitoring, neuroscience, and many others have consistently trended toward miniaturization of sensing devices over the past few decades. This miniaturization has enabled new applications in areas such as connected devices, wearables, implantable medical devices, in vivo microscopy and micro-robotics. By integrating visual sensing capabilities in such devices, we can enable improved spatial, temporal, and contextual information collection for artificial intelligence and data processing.

What is limiting the miniaturization of vision systems? – cameras, an integral part of vision systems, are reaching their limits in size reduction due to restrictions of traditional lensed-optics. Standard smaller lens-system camera modules have a thickness of about 10 mm or higher. Reducing the size of lenses lead to a smaller aperture, poorer light collection, and noisy and worse resolution. I achieve an ultra-compact factor of 10x smaller thicknesses and weights by replacing the focusing lens optics with thinner alternate optics and employ a computational approach to perform “focusing” of images.

The key design principle is to place an optimized flat diffractive mask at a very close distance (range of 100s of microns to a couple of millimeters) from the imaging
sensor and employ computational techniques to recover scene information. Such flat camera geometry calls for new ways of modeling the system, methods to optimize the mask design, and developing computational algorithms to recover desired outputs. With the developed methods, I demonstrate applications in (1) 3D microscopy, (2) high-resolution imaging, (3) low-level vision of oriented-edges feature extraction, and (4) high-level vision of head-pose classification.

With my thesis, I hope to lay the groundwork for creating miniature computational cameras that will push the boundaries of machine vision systems and impact a wide variety of applications like IoT, wearables, implantable sensors, in vivo microscopy, micro-robotics and many more.
Acknowledgments

Foremost, I want to express my gratitude to my advisor, Ashok Veeraraghavan, for his guidance, support, and the opportunities he provided. His relentless persuasion to strive for the best has helped me mature in my professional life. I want to thank my co-advisor, Jacob Robinson, for his continuous support in my research. It was the combined support of Ashok and Jacob that has made possible the research presented in this thesis. I want to thank Richard Baraniuk for spawning the research topic of lensless imaging at Rice University. I want thank my committee member Anastasios Kyrillidis for his time and discussions.

I want to thank my friend and collaborator Jesse Adams. It was an immense pleasure to work with him on all the work presented in this thesis. I want to thank Jason Holloway for helping me getting set in the lab in my first few years of Ph.D. I want to thank all the lab members for the fruitful discussions, cheerful chat, and fun group meetups at Valhalla. I am honored to have worked closely with many of them: Jason Holloway, Salil Tambe, Adam Samaniego, Kaushik Mitra, Huaijin Chen, Yicheng Wu, Adithya Pediredla, Ewa Nowara, Mayank Kumar, Manoj Kumar, Jasper Tan, Akshat Dave, Jesse Adams, Sudarshan Nagesh, Shizheng Zhang.

I want to thank all my Indian friends at Rice. Specifically, Mohit Kumar, Abhinav Achreja, Mayank Kumar, Nithya Renganathan, Sharan Naribole, Artee Bansal, and Arun Mahadevan. It was their love and support that helped me get started in the foreign land. Time spent with them is one of my most cherished moments during my Ph.D.

Last, but not the least, I want to thank the most important people in my life: my soon to be fiancé and wife Kshipra Kapoor, my brother Lokesh Boominathan, and my
parents M. Boominathan and B. Rajeswari. Kshipra’s support towards the end of my Ph.D. is invaluable and it was cheerfulness that kept me sane during defense crunch times. My parents uncoditional love, support, and prayers is what has protected me throughout my life. Finally, I want to thank the Lord and Jesus Christ for His grace and mercy.
Contents

Abstract ii
List of Illustrations x

1 Introduction 1
  1.1 Organization ................................................. 3
  1.2 Disclaimer ..................................................... 4

2 Background: Lensless imaging systems 6
  2.1 Early lensless imaging systems ................................. 8
    2.1.1 Pinhole cameras ........................................... 8
    2.1.2 Coded aperture cameras .................................. 9
    2.1.3 Zone plates ............................................... 10
  2.2 Contemporary lensless imaging systems ......................... 11
    2.2.1 Programmable apertures .................................. 11
  2.3 Diffraction gratings ............................................ 12
  2.4 Shadow and diffraction imaging ............................... 13
    2.4.1 Diffuser caustics ......................................... 14
    2.4.2 Fresnel zone aperture ..................................... 14
  2.5 A mathematical model for lensless imaging ..................... 14
    2.5.1 Lensless imaging architecture ............................ 14
    2.5.2 Image formation .......................................... 15
    2.5.3 Image reconstruction ...................................... 16
  2.6 FlatCam: A lensless imaging case study ....................... 18
    2.6.1 Architectural overview .................................... 18
3 Ultraminiature Fluorescence Microscope

3.1 Introduction .............................................. 29
3.2 Imaging model ........................................... 33
    3.2.1 Amplitude mask design ............................ 33
    3.2.2 Texas Two-Step model ............................ 34
    3.2.3 Model complexity analysis ...................... 37
    3.2.4 Model calibration ................................. 40
3.3 Reconstruction algorithms .............................. 42
    3.3.1 Gradient direction for the T2S model ........... 42
3.4 Prototype ................................................. 43
    3.4.1 Imaging sensor and mask ........................ 43
    3.4.2 Fabrication ...................................... 44
    3.4.3 Calibration ...................................... 44
3.5 Experimental results .................................... 46
    3.5.1 Lateral resolution ................................. 46
    3.5.2 Computational depth scanning ................... 48
    3.5.3 2D Fluorescence imaging ....................... 48
    3.5.4 3D volume reconstruction ...................... 49
    3.5.5 3D volumetric video ............................. 51
3.6 Discussion ................................................ 54

Appendices .................................................. 56
3.A Comparison with FlatCam model ..................... 56
3.B Simulation comparison ........................................ 56
3.C Refractive index matching ..................................... 59
3.D Aberration removal before reconstruction .................. 60
3.E Filter autofluorescence compensation ........................ 60
3.F Diffraction effects on T2S model ............................. 63
3.G Impact of sensor saturation .................................... 63

4 Light efﬁcient high-resolution lensless imaging .......... 66
4.1 Introduction ..................................................... 66
4.2 Diffractive masks ............................................... 69
  4.2.1 Amplitude Masks ............................................ 69
  4.2.2 Phase Masks ................................................ 70
4.3 Diffractive Camera Model ...................................... 71
  4.3.1 Imaging Architecture ...................................... 71
  4.3.2 Fresnel Propagation ........................................ 72
  4.3.3 Point Spread Function ..................................... 72
  4.3.4 Convolutional model ...................................... 74
4.4 Diffractive mask design ........................................ 75
4.5 PSF Engineering ................................................ 78
  4.5.1 Previous lensless imaging PSFs .......................... 78
  4.5.2 Performance and PSF ...................................... 79
4.6 Proposed Contour-based PSF ................................. 81
4.7 Prototype ....................................................... 82
4.8 Reconstruction Algorithm ...................................... 83
4.9 Experimental Results .......................................... 84
  4.9.1 Resolution ..................................................... 84
  4.9.2 Photography ................................................ 86
  4.9.3 Fluorescence Microscopy ................................. 86
5 Miniature machine vision systems

5.1 Introduction ......................................................... 89
5.2 ASOCS design paradigm ........................................... 91
  5.2.1 Step 1: Application-dependent image feature selection .... 94
  5.2.2 Step 2: Phase mask optimization ............................. 95
  5.2.3 Step 3: Fabrication ........................................... 95
5.3 Applications .......................................................... 96
  5.3.1 High resolution imaging ........................................ 96
  5.3.2 Oriented edge extraction ...................................... 97
  5.3.3 Head pose classification ....................................... 100
  5.3.4 Discussion ..................................................... 102

6 Conclusion ................................................................. 104

Bibliography ............................................................... 106
1.1 Outline of demonstrated applications and corresponding chapters. . . 2

2.1 The process of capturing an image using a lensless camera. An additional step of computation is required to reconstruct a clear image from the muddled sensor data. A few examples of lensless cameras [1, 2, 3, 4] are also shown. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

2.2 Schematic of a lensless imager using a single amplitude mask. . . . . . 15

2.3 FlatCam architecture. Every light source within the camera field-of-view contributes to every pixel in the multiplexed image formed on the sensor. A computational algorithm reconstructs the image of the scene. (Left) Inset shows the mask-sensor assembly of our prototype, in which a binary, coded mask is placed 1.2 mm away from an off-the-shelf digital image sensor. (Middle) An example of sensor measurements. (Right) Image reconstructed by solving a computational inverse problem of the form Eq.(2.3). Figure modified from [4]. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19

2.4 Calibration for measuring the left and right matrices $\Phi_L$ and $\Phi_R$ corresponding to a separable mask. (Top) Separable patterns displayed on a screen in front of the camera. The patterns are orthogonal, one-dimensional Hadamard codes that are repeated along either the horizontal or vertical direction. (Bottom) Estimated left and right matrices. Figure modified from [4]. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 21
2.5 Flatcam prototype comparison. Top left is our first prototype with chrome mask placed directly in front of the Flea3 sensor; bottom left is our second prototype with the Omnivision sensor directly epoxied to mask (insets show close-up of the sensors and masks). To the right are reconstruction and BM3D denoised reconstructions for each prototype. The smaller feature size and pixel pitch of the Omnivision prototype provide superior resolution at the cost of more noisier image reconstruction.

2.6 (Top) Reconstruction of three static scenes using the Flea FlatCam prototype. (Bottom) Sample frames from video reconstruction of a toy performing a backflip aided by human hands. The video was recorded at 60 frames-per-second.

2.7 3D imaging with FlatCam. (Center) Experimental setup showing “FLAT” and “CAM” at different distances from the camera. (Left) Heat map of reconstruction PSNR of a simulation of the scene as a function of the scene distance and the calibration distance of the measurement matrices. At closer scene distances, the reconstruction is sensitive to the choice of multiplexing matrix at the correct calibration depth; at further scene distances, the sensitivity decreases. (Right) Reconstruction at 7 cm and 27 cm through simulation and our prototype FlatCam. The word “FLAT” placed at 7 cm is in focus when reconstructed using measurement matrices calibrated to depth of 7 cm. The word “CAM” placed at 27 cm is in focus when reconstructed using measurement matrices calibrated to depth of 27 cm.
3.1 Traditional microscope vs. FlatScope. (A) Traditional microscopes capture the scene through an objective and tube lens (∼20 – 460 mm) resulting in a quality image directly on the imaging sensor. (B) FlatScope captures the scene through an amplitude mask and spacer (∼0.2 mm) and computationally reconstructs the image. Scale bar, 100 µm (inset, 50 µm). (C) Comparison of form factor and resolution for traditional lensed research microscopes, GRIN lens microscope, and FlatScope. FlatScope achieves high-resolution imaging while maintaining a large ratio of field of view (FOV) relative to the cross-sectional area of the device. (D) FlatScope prototype (shown without absorptive filter). Scale bars, 100 µm.

3.2 T2S derivation. (A) Illustration of scene, mask and sensor positions. (B) Sensor pixel’s 2D acceptance profile $C(u,v)$, which can be written as an outer product of the pixel’s response along rows ($c_1(u)$) and the pixel’s response along columns ($c_2(v)$).

3.3 Texas Two-Step model. (A) Illustration of the FlatScope model using a single fluorescent bead as the scene. (B) Fluorescent point source at depth $d$, represented as input image $X_d$. (C) FlatScope measurement $Y$. FlatScope measurement can be decomposed as a superposition of two patterns, (D) pattern when there is no mask in place (open) and (E) pattern due to the coding of the mask. Each of the patterns are separable along x- and y-directions and can be written as (F and G) two separable transfer functions. The FlatScope model, which we call as Texas Two-Step model (T2S), is the superposition of the two separable transfer functions.
3.4 Calibration setup. (A) Side view of the calibration setup (not to scale) showing the LED array, wide angle diffuser (\(~10\) cm above the LEDs), target with line slit (\(~1\) cm above the diffuser) and FlatScope (filter, mask, spacer & sensor). (B) Calibration of rows by translating the horizontal slit target along the y-axis. (C) Calibration of columns by translating the vertical slit target along the x-axis. . . . . . . . . . 41

3.5 Fabrication of FlatScope. Steps for fabricating the amplitude mask and spacer, aligning and fixing to the imaging sensor, adding the absorptive filter and insulating/protecting the device. . . . . . . . . . 45

3.6 Lateral Resolution tests with double slits (A)Double slit with a 1.6 \(\mu\)m gap imaged with a 10\(\times\) objective. (B) Captured FlatScope image. (C) FlatScope reconstruction of the double slit with a 1.6 \(\mu\)m gap. . . 47

3.7 Computational depth scanning with USAF target. FlatScope reconstructions of USAF resolution target at distances from the mask surface of 200 \(\mu\)m (A), 525 \(\mu\)m (B) and 1025 \(\mu\)m (C). Scale bar, 100 \(\mu\)m. 49

3.8 2D Fluorescence imaging of 10 \(\mu\)m fluorescent beads. (A) FlatScope reconstruction. (B) Reference image using an epifluorescence microscope used as ground truth. Scale bar, 100 \(\mu\)m (inset, 50 \(\mu\)m). . 50
3.9 3D Volume Reconstruction of 10 µm fluorescent beads suspended in agarose. (A) FlatScope reconstruction is shown as a maximum intensity projection along the z-axis as well as a ZY slice (blue box) and an XZ slice (red box). (B) Estimated 3D positions of beads from the FlatScope reconstruction. (C) Reference data from a confocal microscope used as ground truth. (D) The depth profile of reconstructed beads compared to ground truth confocal images. Empirically, we can see that the axial spread of 10 µm beads is around 15 µm in FlatScope reconstruction. That is, FlatScope’s depth resolution is less than 15 µm. The three beads shown are at depths of 255 µm, 270 µm and 310 µm from the top surface (filter) of the FlatScope.

3.10 3D Volumetric Video Reconstruction of moving 10 µm fluorescent beads. (A) Sub-section of FlatScope time-lapse reconstruction of 3D volume with 10 µm beads flowing in microfluidic channels (approximate location of channels drawn to highlight bead path and depth). Scale bar, 50 µm (FlatScope prototype graphic at top not to scale). (B) Captured images of frames 1, 3 and 5 (false colored to match time progression). (C and D) FlatScope reconstructions of frames 1, 3 and 5 at estimated depths of 265 µm and 355 µm, respectively (dashed lines indicate approximate location of microfluidic channels). Beads and reconstruction false colored to match time progression. Scale bar, 50 µm.
### 3.B.1 Simulation comparison of bare sensor, FlatScope, and FlatCam.

Reconstructions are shown at depths 0.083 mm, 1.5 mm and 2.5 mm from respective systems. (A) Reconstruction of bare sensor decays rapidly with increasing depth. Reconstruction of FlatScope shows high resolution and stays fairly stable with increasing depth. (C) FlatCam is unable to reconstruct at smaller depths due to limitations of the model with a single separable term. On the other hand, the T2S model of FlatScope can handle all depths. At larger depths, FlatCam and FlatScope reconstructions are comparable.

### 3.C.1 Refractive index matching.
Comparison of reconstructed images captured with different media between the surface of the mask and the target, (A) immersion oil, (B) $H_2O$, and (C) air. Images were captured at a depth of 200 $\mu$m. Note that all images for the calibration process were captured with immersion oil between the surface of the mask and the target. Scale bar, 100 $\mu$m.

### 3.D.1 Aberration removal using Robust PCA.
(A) Captured FlatScope image before aberration removal. (B) Processed capture without aberrations. (C) Aberrations removed by using Robust PCA. (D) Reconstructed image prior to Robust PCA shows large artifacts. (E) Processed reconstruction with aberration removal results in substantially reduced artifacts. (F) Ground truth image. Scale bars, 50 $\mu$m.

### 3.E.1 Removing autofluorescence using DCT.
(A) Captured FlatScope image before DCT removal. (B) Processed capture image. (C) Reconstructed image prior to DCT removal shows excessive noise. (D) Processed reconstruction with DCT removal results in substantially reduced noise. Scale bars, 100 $\mu$m.
3. F. 1 T2S model error from diffraction. (A) Raw capture by FlatScope prototype of a point source placed at a depth of 250 µm. The raw capture is cropped to the region of interest. (B) T2S model approximation of the capture. (C) Mask region illuminated by the point source. The location of the point source was chosen such that the illuminated mask region has a diversity in aperture sizes (white represents open apertures; black represents closed apertures). (D) Absolute error difference between raw capture and T2S model approximation. Average error is 12.64%. Scale bar, 40 µm.

1 An outline of mask design and system architecture for our miniature diffractive camera. We first obtain a target point spread function (PSF) through PSF engineering. Then, using our iterative phase mask design, we optimize for a phase map that can reliably produce the target PSF.

2 Target and achieved PSF using (a) binary amplitude mask and (b) phase mask. The PSF from amplitude mask is heavily degraded by diffraction effects.

3 Forward Propagation.

4 Phase mask optimization.
5 Analysis of various lensless point-spread-functions (PSF). (a) Separable PSF created from outer product of two m-sequences as used by [4]. (b) Fresnel zone apertures (FZA) as used by [5]. (c) Tessellated spiral PSF as used by [6]. A single unit of the tessellation is shown on the top left corner. Tessellation is needed to cover a larger sensor area. (d) Our proposed Perlin noise based PSF. (e) Magnitude spectrum of each PSF. The magnitude spectrum of Contour PSF remains large for entire frequency range indicating better invertibility characteristic. (f) Singular values of convolution matrix for each PSF. Singular value of Contour PSF remains fairly larger than for other PSFs indicating high performance. The large initial values for the spiral PSF is due to frequency harmonics caused by tessellation. (g) Mean Squared Error (MSE), normalized over noise variance, of reconstruction gives performance measure in the presence of noise. Our proposed Contour PSF has low MSE compared to other PSFs indicating high performance.

6 Generating Contour PSF

7 Comparison with caustic pattern produced by 0.5° diffuser. Highest contrast pattern with diffuser was achieved at 16 mm. Our proposed Contour PSF has lower MSE compared to diffuser caustic and hence will perform better.

8 (a) Phase mask design. (b) Comparison between target PSF and captured PSF using the designed and implemented phase mask.

9 Geometry of lensless camera. Theoretical resolution can be calculated from the similar triangles formed between pixel pitch and smallest achievable feature resolution.
10 Experimental evaluation of our camera’s resolution using fluorescent USAF target. The inserts are shown for line pairs with contrast close to 20% or higher. ................................. 85

11 Photography experiment. (a) Photography setup. (b) Capture and reconstruction. ................................. 86

12 Microscopy experiment. (a) Microscopy setup. (b) Capture and reconstruction. ................................. 87

1 Miniature Computer Vision System Design. (a) Traditional machine vision system has a general purpose optical imaging system followed by application-specific computation to achieve an application-specific goal. (b) In our miniature machine vision system, we design a front end application-specific optical computing systems (ASOCS) with thickness constraint followed by computational algorithms to achieve an application-specific goal. In this paper, we explore three different types of applications: imaging, oriented edge-extraction, and head pose classification. ................................. 92

2 ASOCS Design and Example applications. (a) Considering the application-specific goals and constraints, we design and develop the processes outlined in dashed lines. (b) Table laying out the goals, constraints, and ASOCS design components for three different applications: edge-extraction, and head pose classification. ................................. 93

3 High-resolution imaging. (a) Comparison of traditional vs our miniature imaging system. The ASOCS was designed with thickness constraint of 1.95 mm. (b) Proposed contour PSF. (c) Optimized phase mask profile for the proposed contour PSF. ................................. 97
4 Miniature oriented-edge extraction system (a) Comparison of traditional system of extracting edge features vs our miniature system. The ASOCS was designed with thickness constraint of 4 mm. (b) Positive and negative components of Gabor filters of orientations 90deg and 45deg. (c) Optimized phase masks to implement the components of Gabor filters.

5 Oriented-edge extraction experiments. (a) Experimentally realized filters and edge extractions at orientations 45deg and 90deg. The house image was captured by displaying on a monitor display while the face was captured directly. (b) Various filters that can be realized from the implemented PSFs using per-pixel subtraction.

6 Head pose classification. The ASOCS design was learned for thickness constraint of 1.95 mm. (b) Learned PSFs from CNN split into positive and negative components. (c) Optimized phase masks to realize the split PSFs. (d) Classification performance from implemented prototype.

7 CNN architecture and performance. (a) CNN architecture used for training the optical filters and testing performance. (b) Learned PSFs from CNN split into positive and negative components. (c) Optimized phase masks to realize the split PSFs. (d) Classification performance from implemented prototype.
Chapter 1

Introduction

Vision can be attributed as one of the key senses that humans and other animals use to perceive the world. With decades of innovation, imaging technology has evolved to equip machines and electronic devices with the ability to “see” by recording, understanding, and analyzing the world. This paves the way for machine vision systems that is revolutionizing the fields of robotics, health-monitoring, neuroscience, and many others.

Miniaturization is a critical need for machine vision systems in many emerging applications such as wearable and implantable electronics, robotics and automation, and the internet of things (IoT) [8, 9, 10]. Fortunately, advances in cloud computing and Moore’s law [11] have produced dramatic miniaturization of the electronic components necessary for machine vision systems; however, the corresponding optical elements have not achieved comparable size reductions, limiting the scale and scope of adoption in such applications. We note that the size of cameras, crucial part of machine vision systems, is what is limiting the advancement of miniaturization.

The traditional design of cameras uses a lens to focus light from the scene under view onto a photosensitive surface. Over the years, the photosensitive surface has evolved from a photographic film to an array of digital sensors. However, lenses remain an integral part of modern imaging systems and also introduce size constraints. While image sensors are typically thin, cameras end up being thick due to the lens complexity and the large distance required between the lens and sensor to achieve
Figure 1.1 : Outline of demonstrated applications and corresponding chapters.

focus. For example, the thinnest mobile cameras today are approximately 5-mm thick, with the thickness increasing at larger lens aperture sizes [12]. Hence, advances in lens design alone cannot miniaturize cameras. We need an additional tool to flex the design space of camera and achieve miniaturization. Computation could be that tool.

Last few decades has seen an exponential growth in computing power packed into miniature chip volumes due to advances in transistor fabrication technologies [11, 13]. This availability of computing power provides a new avenue for solving traditional imaging problems. Notably, recent advances in computational imaging have used computation to drastically shift the burden from optics to algorithms. Computation allows for significantly reduced optical complexity and size while maintaining high image fidelity. In this thesis work, we use computational tools not to supplement but complement optics design and jointly achieve miniaturization.
Advances in fabrication technologies have made it possible to explore non-smooth optics while lowering the cost of bulk manufacturing of the same. Such non-smooth unconventional optics can be ultra-thin, lightweight and have large and flexible design space. In particular, a class of optics called diffractive optical elements (DOEs) has, recently, drawn considerable attention in the computational imaging community. Focusing DOEs have been used to substitute compound lens system and create lightweight camera designs [14, 15]. In this thesis work, we design and place DOEs at a much smaller distance [16, 17] from an imaging sensor to achieve cameras that are both lightweight and ultra-compact.

The class of imaging systems that replace the lens with diffractive mask and computation is called lensless imaging [4, 18, 19, 17, 20, 21, 22, 23, 7]. In my thesis, I design lensless systems that are 5-10x smaller in thickness and 20x lighter weight than traditional lensed systems and I demonstrate these systems for following applications: (1) 3D microscopy, (2) high-resolution imaging, (3) low-level vision of oriented-edges feature extraction, and (4) high-level vision of head-pose classification. Outline of all the demonstrated applications is shown in Figure 1.1.

1.1 Organization

The remainder of this dissertation will be structured as follows.

Chapter 2: Background and Related Work. This chapter surveys class of miniature cameras called lensless cameras.

Chapter 3: Ultraminiature Fluorescence Microscopy. This chapter describes our work in designing and fabricating ultra-miniature camera for a very specific application of fluorescence microscopy.

Chapter 4: Light efﬁcient high-resolution lensless imaging. This chapter describes
our work in designing high-resolution lensless system with high-efficient phase mask.

Chapter 5: Miniature machine vision systems This chapter takes a step back and lays out a generalized framework for designing light-efficient miniature cameras with target application in mind. The target application is not limited only to imaging but can be used for designing miniature machine vision systems.

Chapter 5: Conclusion. This chapter summarizes the thesis contributions.

1.2 Disclaimer

The results in this thesis, and in large part, the text of this thesis is adapted from:


as well as a manuscript currently in preparation:

Vivek Boominathan, Jesse K. Adams, Jacob T. Robinson, Ashok Veeraraghavan, “Miniature machine vision systems via application-specific optical com-
puting”, *In preparation*
Chapter 2

Background: Lensless imaging systems

In this chapter, we review a variety of alternate imaging approaches that completely eschew lenses. The primary task of a lens in a camera is to shape the incoming light wavefront so that it creates a focused image on the sensor. In the absence of a lens, a sensor would simply record the average light intensity from the entire scene. *Lensless imaging systems* dispense with a lens by using other optical elements to manipulate the incoming light. The sensor records the intensity of the manipulated light, which may not appear as a focused image. However, when the system is designed correctly, the image can be recovered from the sensor measurements with the help of a computational algorithm. Figure 2.1 shows the processes for capturing/reconstructing images in lensed and lensless systems. The simplest lensless imaging system is the pinhole camera. It is inefficient, however, since the small pinhole restricts the amount of light reaching the sensor. Coded aperture cameras improve the light efficiency using a mask with an array of pinholes. The sensor measurements become a superposition of the images formed by each aperture, and the computational recovery algorithm’s task is to reorganize the measurements to recover the image.

There are many benefits to going lensless:

- *Scalable fabrication.* Lensless cameras can be directly fabricated using traditional semiconductor fabrication technology. For example, a multiple-aperture mask can be fabricated either directly in one of the metal interconnect layers
Figure 2.1: The process of capturing an image using a lensless camera. An additional step of computation is required to reconstruct a clear image from the muddled sensor data. A few examples of lensless cameras [1, 2, 3, 4] are also shown.

or on a separate wafer thermal compression that is bonded to the back side of the sensor, as is typical for back-side-illuminated image sensors [24]. Thus, lensless cameras can benefit from all of the scaling advantages of semiconductor fabrication, resulting in a low-cost, high-yield, high-performance device. In contrast, conventional cameras require inefficient post-fabrication assembly of the lens system.

- **Thin form factor.** Since the standoff distance between a multiple-aperture mask and the image sensor array need only be a few tens to hundreds of microns, an entire lensless camera can be only a few tens to hundreds of microns thick—resulting in potentially the thinnest cameras ever produced.

- **Non-planar geometries.** Lensless cameras can be adapted to arbitrary sensor geometries, including not just planar but also cylindrical, spherical, and even
flexible sensors. The compact form of spherical lensless cameras promise unmatched maneuverability in constrained environments such as endoscopy.

- **Light throughput.** Lensless cameras can be designed to have very large input apertures, which translates into improved light efficiency and much larger field-of-view than conventional lens-based systems.

- **Three-dimensional (3D) imaging.** Lensless imaging systems can extract 3D and refocusing information in addition to 2D imaging. Although this ability is not yet competitive with existing lens based techniques such as light-field and time of flight, the extracted 3D information may still be useful in some contexts such as gesture identification.

In this chapter, we review the past, present, and future of lensless imaging as a shining example of the opportunities afforded by computational imaging, a design framework that uses computational algorithms to replace or augment imaging hardware (in this case replacing the lens). After reviewing classical and contemporary approaches to lensless imaging, we introduce and analyze a mathematical model that exposes the key issues underlying these architectures. The bulk of the paper consists of a case study of the FlatCam [4], a particular mask-based lensless imager we have developed.

### 2.1 Early lensless imaging systems

#### 2.1.1 Pinhole cameras

The very first cameras were lensless. Pinhole cameras, also known as the camera obscura, were discovered centuries before the invention of lenses and photography.
Pinhole cameras have been well known since Alhazen (965-1039 AD) and Mozi (c. 370 BCE). However, the first photograph using a pinhole camera was captured in 1850. Pinhole cameras offer a simple and elegant architecture for lensless imaging that consists of a single aperture in front of a sensor. Light from an object passes through the pinhole and forms its image on the sensor. However, a tiny pinhole is required to produce sharp images, which results in very low light throughput. As a consequence, a pinhole camera requires very long exposure times to acquire images at high quality. Indeed, lenses were introduced into cameras for precisely the purpose of increasing the size of the aperture, and thus the light throughput, without degrading the sharpness of the acquired image.

2.1.2 Coded aperture cameras

Coded aperture cameras extend the idea of a pinhole camera by replacing the small, single aperture with a mask containing multiple apertures [25, 26, 27]. Coded aperture cameras were originally invented for imaging with X-rays and gamma-rays, wavelengths of light that are not easily amenable to lens-based imaging. In a general coded aperture system, sensor measurements represent a superposition of the images formed behind each pinhole. The primary motivation for a coded mask is to increase the light throughput while retaining the ability to reconstruct high-resolution images. For instance, if the mask contains \( P \) pinholes, then the sensor image is the sum of \( P \) overlapping images of the scene. The signal-to-noise ratio in such an image is approximately \( \sqrt{P} \) times better than a single pinhole image [25, 26].

In contrast to a single-pinhole camera, the sensor measurements of a coded aperture camera do not resemble an image of the scene. Rather, each light source in the scene casts a unique shadow of the mask onto the sensor, encoding information about
locations and intensities. Consider a single light source on a dark background; the image formed on the sensor will be a shadow of the mask. If we change the angle of the light source, then the mask shadow on the sensor will shift. If we change the depth of the light source, then the size of the shadow will change. We can represent the relationship between the scene and the sensor measurements as a linear system that depends on the pattern and placement of the mask. Inverting this system using an appropriate computational algorithm will recover an image of the scene.

The design of the mask plays an important role in coded-aperture imaging. An ideal pattern would maximize the light throughput while providing a well-conditioned scene-to-sensor transfer function to facilitate inversion. In this regard, several mask designs have been proposed in coded aperture literature, including Fresnel zone plate, random pinhole patterns, uniformly redundant arrays (URAs) [26], and their extensions. URAs are particularly useful because of two key properties: (1) almost half of the mask is open, which boosts the signal-to-noise ratio; (2) the autocorrelation function of the mask is close to a delta function, which aids in calibration and image recovery. URA patterns are closely related to the Hadamard-Walsh functions and the maximum length sequences that are maximally incoherent with their cyclic shifts [28].

2.1.3 Zone plates

A zone plate can also be used to focus light and form an image using diffraction [29, 30]. A zone plate consists of concentric transparent and opaque rings (or zones). Light hitting a zone plate diffracts around the opaque regions and interferes constructively at the focal point. Zone plates can be used in place of pinholes or lenses to form an image. One advantage of zone plates over pinholes is their large transparent area,
which provides better light efficiency. In contrast with lenses, zone plates can be used for imaging wavelengths where lenses are either expensive or difficult to manufacture [31, 32].

2.2 Contemporary lensless imaging systems

Recent advances in sensor technology (in particular, the conversion from analog film to digital CCD and CMOS sensor arrays), image reconstruction models and algorithms, and computing resources have made lensless imaging a burgeoning field. Here we briefly review some of the recent research in this area.

2.2.1 Programmable apertures

Programmable mask-based lensless imaging designs have recently been proposed in [33, 1, 34]. The camera proposed in [33] consists of a sensor and layers of programmable spatial light modulators (SLMs) whose transmittances are controllable in space and time. By applying different patterns in each layer, the incoming light can be manipulated in a number of ways. For example, the camera can track a moving object by shifting a pinhole in one of the layers, select and capture disjoint regions in the scene, or perform computations on the scene and record the results directly on the sensor.

The lensless camera in [1] (first example of lensless in Figure 2.1) uses compressive sensing principles to capture and recover images. It consists of a single programmable SLM and a single pixel detector. It captures multiple measurements of the scene by changing the mask pattern. The scene is then reconstructed by solving a sparse recovery program. Using multiple pixel detectors, this design can reconstruct a higher resolution image for a planar or a sufficiently distant scene [35].
The camera in [34] consists of a sensor array and an SLM implementing a separable mask pattern. This camera can reconstruct the scene using a single sensor image, but the reconstruction quality improves using multiple sensor images with different mask patterns. In the development of this camera, the authors showed that traditional techniques [36] of using URA and modified URA (MURA) aperture patterns fail due to significant diffraction effects in the visible spectrum.

2.3 Diffraction gratings

Ultra-miniature cameras (approximately 100 µm width and thickness) have been implemented in [22, 37, 2, 7] using integrated diffraction gratings and CMOS image sensors. The pixels in [22] use diffraction gratings over a photodiode in order to be sensitive to the angle of incident light. The angle selectivity is achieved due to a phenomenon called the Talbot effect [38] and enables the camera to perform lensless 3D imaging in the near-field. The gratings were fabricated as metal wiring layers over the photodiodes.

The phase gratings in [2] are designed such that they impose spiral-shaped diffraction patterns (second example for lensless in Figure 2.1) on the sensor array. The diffraction pattern is etched on a refractive medium placed above the sensor. The spiral pattern can also be viewed as the point spread function of these imaging systems. Similar to a coded aperture system, the image formed on the sensor is a superposition of shifted and scaled spiral patterns. However, in contrast to an amplitude mask, a phase grating-based mask has improved light efficiency, since it blocks much less light. While an image of the scene can be recovered using a computational algorithm, the primary purpose of these small-size and low-cost designs is distributed monitoring and inspection (for example, in the internet-of-things).
2.4 Shadow and diffraction imaging

Lensless cameras have also been successfully demonstrated for several microscopy and lab-on-chip applications. We can divide the lensless microscopes into two broad categories: contact-mode shadow imaging-based microscopes [39, 40, 41] and diffraction-based lensless microscopes [42, 43, 3, 44, 45, 46]. In a shadow imaging-based microscope, a microscopic sample is placed extremely close to a sensor array (ideally within 1 \( \mu \text{m} \)) so that diffraction is minimized. Light from an illumination source passes through the sample and casts a shadow on the sensor with unit magnification. The shadow image represents the image of the microscopic sample under observation. It is also possible to capture multiple images of a sample with subpixel shifts for the purpose of digital superresolution. The on-chip microscope in [40] demonstrated imaging of red blood cells at a resolution of 600 nm by combining multiple low-resolution shadow images of blood flowing in a microfluidic channel.

Diffraction-based lensless microscopes allow a significant distance between the sample and the sensor plane. Light scattered by the sample interferes with itself and creates an interference pattern on the sensor (third example of lensless in Figure 2.1). These interference patterns can be digitally processed to reconstruct an image of the sample [3, 44]. The on-chip microscope in [44] demonstrated imaging of red blood cells at a resolution less than 7 \( \mu \text{m} \) with a field-of-view of 20.5 mm\(^2\). Since the optical sensor records only the intensity of the interference patterns and loses the phase information, image reconstruction relies on computational methods for phase retrieval [47, 48].
2.4.1 Diffuser caustics

An inexpensive lensless imaging system with diffuser, like scotch tape, as an encoding mask was demonstrated for 3D imaging \cite{18} and 3D fluorescence microscopy \cite{49}.

2.4.2 Fresnel zone aperture

A retake on the Fresnel zones, but, instead placed closer to the sensor to behave like a coded mask was used to demonstrate lensless imaging for a rudimentary light-field imaging \cite{19} and close-up imaging \cite{20}.

2.5 A mathematical model for lensless imaging

A simple mathematical model can be used to explain, characterize, and analyze the operation of a variety of lensless imagers.

2.5.1 Lensless imaging architecture

Consider the imaging architecture in Figure \ref{fig2.2} which consists of an amplitude mask placed in front of an image sensor. Both the sensor and the mask are assumed to be planar and parallel to each other. The mask is placed a distance $d$ (typically measured in microns) in front of the sensor; hence, we can assume the sensor is placed on the plane $z = 0$ and the mask on the plane $z = d$. Assume without loss of generality that the mask is binary-valued and consists of opaque and transparent elements that either block or transmit light. An important variable is the smallest feature size on the mask, $\Delta$; intuitively, the binary mask is constructed using opaque or transparent building blocks of size $\Delta \times \Delta$. Denote the pixel pitch, or the size of individual sensor pixels, by $w$. Given this basic setup, we can characterize the spot size produced
by a mask element and characterize when the spot can be well-approximated using geometric (ray) optics.

2.5.2 Image formation

We characterize image formation using the geometric optics model. While this approach largely ignores diffraction, the resulting model is useful for the design and analysis of well-conditioned imaging architectures. Furthermore, the calibration procedure that we detail in subsequent sections can account for un-modeled diffraction effects. For the simplicity of notation, we assume a simplified 2D world imaged by a 1D mask and sensor. The extension to a 3D world imaged by a 2D mask and sensor is straightforward except where stated otherwise.

For a suitably defined scene irradiance vector $\mathbf{x} \in \mathbb{R}^N$, the scene-to-sensor mapping can be described using the linear set of equations

$$\mathbf{y} = \Phi \mathbf{x} + \mathbf{e}$$  \hspace{1cm} (2.1)

where $\Phi \in \mathbb{R}^{M \times N}$ is the measurement matrix, $\mathbf{y} \in \mathbb{R}^M$ is the image formed on the
sensor, and \( e \) is measurement noise. This model can be interpreted in two different ways: (1) Each sensor measures a weighted, linear combination of light from multiple scene locations, and each row in \( \Phi \) encodes the weights for the respective sensor. For a scene at infinity, the weights for two different sensor pixels simply differ by a translation of the mask pattern. As a consequence, the matrix \( \Phi \) has a Toeplitz structure. (2) Every light source in the scene casts a shadow of the mask on the sensor. Thus, the image formed on the sensor is a superposition of shifted and scaled versions of the mask. The shift and the scaling of the mask pattern encodes the angle and distance of the light source onto the sensor. These properties are invaluable in the design of masks that provide near-optimal recovery under noise. Given the image formation model in (2.1), our tasks are to formulate an inversion algorithm that recovers the scene \( x \) from the sensed image \( y \) and design mask patterns that achieve optimal recovery performance. We study both problems in the subsequent sections.

2.5.3 Image reconstruction

Given the sensor measurements \( y \in \mathbb{R}^M \) and the measurement matrix \( \Phi \), recovering \( x \in \mathbb{R}^N \) depends mainly on the rank of the matrix \( \Phi \) and its condition number. When \( \text{rank}(\Phi) = N \) and the matrix is well-conditioned, we can obtain an estimate of \( x \) by solving the least-squares problem

\[
\min_x \| \Phi x - y \|_2^2,
\]

which has the closed form solution \( \hat{x}_{LS} = \Phi^+ y = x + \Phi^+ e \), where \( \Phi^+ = (\Phi^T \Phi)^{-1} \Phi^T \) is the pseudoinverse of \( \Phi \). When \( \Phi \) is not well-conditioned, the least squares estimate \( \hat{x}_{LS} \) suffers from noise amplification. When \( \Phi \) is rank-deficient, the matrix becomes singular and an estimate cannot be achieved.
In the ill-conditioned and rank-deficient cases, we can use an image prior to regularize the inverse problem. Specifically, instead of solving (2.2), we solve

$$\min_x \|y - \Phi x\|^2_2 + \lambda \mathcal{R}(x),$$

(2.3)

where $\|y - \Phi x\|^2$ quantifies the data fidelity, $\mathcal{R}(x)$ is a regularization term that enforces an image prior, and $\lambda > 0$ controls the trade off between fidelity and regularization. A popular choice for the regularizer that is useful for noise-suppression is Tikhonov regularization (a.k.a. ridge regression) via $\mathcal{R}(x) = \|x\|^2_2$.

Natural signals such as images and videos exhibit a host of geometric properties including sparse gradients and sparse coefficients in certain transform domains (e.g., Fourier or wavelets). By enforcing these geometric properties, we can suppress noise amplification as well as obtain unique solutions even when $\Phi$ is rank-deficient (i.e, $M < N$). A pertinent example for image reconstruction is the total-variation (TV) model, where the regularizer $\mathcal{R}(x) = \|x\|_{TV}$ corresponds to the TV of the image, which is computed from its gradients. Writing the scene $x$ as the 2D image $x(u, v)$ and defining $g_u = D_u x$ and $g_v = D_v x$ as the $u$- and $v$-components, respectively, of the spatial gradient of the image, the TV of the image is given by

$$\mathcal{R}(x) = \|x\|_{TV} = \sum_{u,v} \sqrt{g_u(u, v)^2 + g_v(u, v)^2}.$$

The minimization (2.3) with a TV prior is convex and produces images with sparse gradients. A host of efficient techniques have been developed to obtain the solution. A range of even more realistic image models have been developed (e.g., [50]), but the resulting optimization might not be convex.
2.6 FlatCam: A lensless imaging case study

To illustrate the design tradeoffs involved in a practical lensless camera design, we review the FlatCam [4], which was inspired by the coded aperture imaging principles pioneered in astronomical X-ray and Gamma-ray imaging [51, 36, 52, 53].

2.6.1 Architectural overview

The FlatCam design achieves a large photosensitive area with a thin form factor by replacing the lens with a coded, binary mask. The thickness of the camera is minimized by placing the mask almost immediately on top of a bare conventional sensor array. The image formed on the sensor can be viewed as a superposition of many pinhole images. An illustration of the FlatCam design is presented in Figure 2.3. Light from all points in the scene passes through a coded mask and forms a multiplexed image on the sensor. A computational algorithm is used to recover the original light distribution of the scene from the sensor measurements.

The FlatCam design has many attractive properties besides its slim profile. First, since it reduces the thickness of the camera but not the area of the sensor, it can collect more light than a miniature, lens-based camera of the same thickness. The light collection ability of FlatCam is proportional to the size of the sensor and the transparent regions (pinholes) in the mask. In contrast, the light collection ability of a lens-based camera is limited by the lens aperture size, which is restricted by the requirements on the device thickness. Second, the mask can be created from inexpensive materials that operate over a broad range of wavelengths. Third, the mask can be fabricated simultaneously with the sensor array, creating new manufacturing efficiencies.
Figure 2.3: FlatCam architecture. Every light source within the camera field-of-view contributes to every pixel in the multiplexed image formed on the sensor. A computational algorithm reconstructs the image of the scene. (Left) Inset shows the mask-sensor assembly of our prototype, in which a binary, coded mask is placed 1.2 mm away from an off-the-shelf digital image sensor. (Middle) An example of sensor measurements. (Right) Image reconstructed by solving a computational inverse problem of the form Eq. (2.3). Figure modified from [4].

2.6.2 Mask design and calibration

Separable masks The FlatCam uses a separable mask pattern, i.e., the 2D mask pattern is the outer product of two 1D patterns. Such a pattern drastically reduces the storage and computational footprint of the measurement matrix $\Phi$. When the mask pattern is separable, the imaging equation (2.1) can be rewritten as

$$Y = \Phi_L X \Phi_R^T + E,$$  \hspace{1cm} (2.4)

where $X$ is an $N \times N$ matrix containing the scene radiance; $Y$ in an $M \times M$ matrix containing the sensor measurements; $\Phi_L$ and $\Phi_R$ are matrices representing 1D convolution along the rows and columns of the scene, respectively; and $E$ denotes the sensor noise and model mismatch. For a megapixel scene/image and a megapixel
sensor, \( \Phi_L \) and \( \Phi_R \) each have only \( 10^6 \) elements, as opposed to \( 10^{12} \) elements in \( \Phi \). A similar idea has been recently proposed in [34] with the design of doubly Toeplitz mask.

**Mask design** The mask pattern should be chosen to make the matrices \( \Phi_L \) and \( \Phi_R \) as numerically stable as possible, which ensures a stable recovery of the image \( X \) from the sensor measurements \( Y \). In the context of image reconstruction using signal priors (for example, TV prior discussed above), random matrices enjoy stable recovery guarantees. Hence, we construct the separable mask pattern as the outer-product of two 1D pseudo-random sequences.

**Calibration** The low-dimensionality of \( \Phi_L \) and \( \Phi_R \) in (2.4) support a simple and efficient calibration scheme. Instead of modeling the convolution shifts and diffraction effects for a particular mask-sensor arrangement, we directly estimate the system matrices from training data. To align the mask and sensor, we adjust their relative orientation such that a separable scene in front of the camera yields a separable image on the sensor. For a perfectly aligned system, displaying a horizontal/vertical line on a screen in front of the camera results in an image containing a set of sharp horizontal/vertical stripes. We first achieve sharpness by rotating the mask relative to the screen. Then, we align the sensor and mask so that the stripes on the sensor image are parallel to the image axis. To calibrate a system that can recover an image \( X \) with dimensions \( N \times N \), we estimate the left and right matrices \( \Phi_L \) and \( \Phi_R \) using the sensor measurements of \( 2N \) known calibration patterns projected on a screen as depicted in Fig. [2.4]. Our calibration procedure relies on an important observation:
Figure 2.4: Calibration for measuring the left and right matrices $\Phi_L$ and $\Phi_R$ corresponding to a separable mask. (Top) Separable patterns displayed on a screen in front of the camera. The patterns are orthogonal, one-dimensional Hadamard codes that are repeated along either the horizontal or vertical direction. (Bottom) Estimated left and right matrices. Figure modified from [4].

If the scene $X$ is separable, i.e., $X = ab^T$ where $a, b \in \mathbb{R}^N$ then, for an ideal system,

$$Y = \Phi_L ab^T \Phi_R^T = (\Phi_L a)(\Phi_R b)^T.$$

In essence, the image formed on the sensor is a rank-1 matrix, and using a truncated singular value decomposition (SVD), we can obtain estimates of $\Phi_L a$ and $\Phi_R b$ up to a signed, scalar constant. We take $N$ separable pattern measurements for calibrating each of $\Phi_L$ and $\Phi_R$. In practice, we average several measurements of each calibration pattern to reduce the effects of sensor noise.
2.6.3 Prototypes

We have built two different FlatCam prototypes. The first prototype consists of a Point Grey Flea3 with a Sapphire EV76C560 CMOS sensor which has a 5.3 µm pixel size and measured Chief Ray Angle (CRA) of 25°. (The CRA of a sensor determines the cone of light that can enter a pixel.) The diffractive mask is chrome on quartz glass placed adjacent to the infrared filter of the sensor (mask-to-sensor distance: 1.2
The pattern on the mask is an outer product of two length-1024 pseudorandom sequences of smallest feature size 25 \( \mu m \).

The second prototype was assembled with a diffractive mask and spacer attached directly to the surface of an Omnivision OV5647 CMOS sensor (fourth example of lensless in Figure 2.1). The Omnivision sensor has pixels of size 1.4 \( \mu m \) and measured CRA of 28\(^{\circ}\). The diffractive mask was fabricated by depositing a thin-film of chrome on fused silica that was then patterned with photoresist and etched to leave the desired pattern. The mask was then diced, aligned to the CMOS pixel array and attached with optical epoxy (mask to sensor distance 500 \( \mu m \)). The pattern on the mask is the outer product of length-1296 and 972 pseudorandom sequences of smallest feature size of 2.8 \( \mu m \). The smaller feature size and pixel pitch of this prototype enable reconstructions at a higher resolution. However, a drawback of the smaller pixel pitch is a sensor with poorer SNR performance which results in noisier measurements and reconstruction compared to our first prototype (see Figure 2.5).

The remainder of the examples were obtained using the higher-quality Flea3 sensor prototype.

2.6.4 Imaging and Video

Sample reconstructions using Flea3 prototype (first prototype) are shown in Figure 2.6 (Top). Reconstructions of a dynamic scene are shown in Figure 2.6 (Bottom); here, we operated the camera with a 3 ms exposure and recovered videos of 60 frames-per-second with each frame of the video recovered independently as a stand-alone image.
Figure 2.6: (Top) Reconstruction of three static scenes using the Flea FlatCam prototype. (Bottom) Sample frames from video reconstruction of a toy performing a backflip aided by human hands. The video was recorded at 60 frames-per-second.

2.6.5 3D imaging

FlatCam can computationally change its focus to new depths in a scene from a single acquisition. The key is that, for a given mask design, we can calibrate a set of separable measurement matrices $\{\Phi_{Li}\}_{i=1,...,L}$ and $\{\Phi_{Rj}\}_{j=1,...,R}$, each obtained using a screen at a different depth (recall Figure 2.4).

Figure 2.7 (left) shows a heat map of reconstruction PSNR of a simulated 2D scene as a function of the scene distance and the calibration distance of the measurement
matrices. We see that the reconstruction quality improves as the calibration depth of the camera approaches the actual scene depth. Moreover, the sensitivity of the reconstruction due to the discrepancy in these depths decreases with increasing scene distance.

Figure 2.7 (right), shows the reconstruction of a 3D scene at two different depth planes. For a particular fixed mask, we calibrated the measurement matrices with a screen at the distances of 7 cm and 27 cm. We accounted for the field of view of the sensor by adjusting the size of the calibration patterns in accordance with the CRA of the sensor. The resulting two sets of matrices were then used to create focused images at 7 cm and 27 cm from a single acquisition with FlatCam. (The line artifacts in the experimental reconstruction are due to scene illumination leaking into the sensor from the sides that was not accounted for in the calibration procedure. We can reduce the unaccounted light in future prototypes by introducing baffles.)

2.7 Limitations and challenges facing lensless imaging

The very first cameras were lensless (pinhole cameras), but the advent of lenses and other advanced optics relegated such systems to niche applications like X-ray and gamma ray imaging. The resurgence of lensless imaging can be attributed to the convergence of four factors: the development of digital CMOS and CCD sensor arrays, efficient and realistic image models and recovery algorithms, powerful computing, and new mask designs (such as the separable mask in the FlatCam).

The further development of lensless imaging, however, will face challenges. As the mask is moved closer to the sensor in any pinhole or coded aperture camera, the angular resolution decreases, resulting in a trade-off between minimal thickness and spatial resolution. Additionally, computationally recovering a scene from less-than-
Figure 2.7: 3D imaging with FlatCam. (Center) Experimental setup showing “FLAT” and “CAM” at different distances from the camera. (Left) Heat map of reconstruction PSNR of a simulation of the scene as a function of the scene distance and the calibration distance of the measurement matrices. At closer scene distances, the reconstruction is sensitive to the choice of multiplexing matrix at the correct calibration depth; at further scene distances, the sensitivity decreases. (Right) Reconstruction at 7 cm and 27 cm through simulation and our prototype FlatCam. The word “FLAT” placed at 7 cm is in focus when reconstructed using measurement matrices calibrated to depth of 7 cm. The word “CAM” placed at 27 cm is in focus when reconstructed using measurement matrices calibrated to depth of 27 cm.
perfectly-conditioned sensor measurements results in noise amplification. Although noise amplification cannot be eliminated, careful design of mask patterns and regularization models can minimize this effect. The necessity for a computational algorithm also results in a time-lag between image acquisition and reconstruction (∼100 ms for FlatCam). Such a delay may be acceptable in certain applications but unacceptable in others such as augmented or virtual reality. There are a number of avenues for continued research and development that could lead to significantly improved lensless imaging performance, including new architectures for improving spatial resolution, new image models to reduce the demultiplexing noise, and new computational algorithms to support high-speed sensing. Sometimes, size matters. The lensless imaging approach promises to challenge the traditional barriers of size, weight, cost, and performance in a broad range of applications spanning consumer, medical, scientific imaging, machine vision, and remote sensing. Indeed, the future of lensless imaging research and development looks very bright.
Chapter 3

Ultramiiniature Fluorescence Microscope

In this chapter, we focus on a specific application of fluorescence microscopy, where miniaturization can have a drastic impact. Miniaturizing traditional microscope architectures suffer from a fundamental tradeoff: as lenses become smaller, they must either collect less light or image a smaller field of view. To break this fundamental tradeoff between device size and performance, we present a new concept for 3D fluorescence imaging that replaces lenses with an optimized amplitude mask placed a few hundred microns above the sensor and an efficient algorithm that can convert a single frame of captured sensor data into high-resolution 3D images. The result is FlatScope: perhaps the world’s tiniest and lightest microscope. FlatScope is a lensless microscope that is scarcely larger than an image sensor (roughly 0.2 grams in weight and less than 1 mm thick) and yet able to produce micron-resolution, high-frame-rate, 3D fluorescence movies covering a total volume of several cubic millimeters. The ability of FlatScope to reconstruct full 3D images from a single frame of captured sensor data allows us to image 3D volumes roughly 40,000 times faster than a laser scanning confocal microscope, and yet providing comparable resolution. We envision that this new flat fluorescence microscopy paradigm will lead to implantable endoscopes that minimize tissue damage, arrays of imagers that cover large areas, and bendable, flexible microscopes that conform to complex topographies.
3.1 Introduction

Revolutionary advances in fabrication technologies for lenses and image sensors have resulted in the remarkable miniaturization of imaging systems and opened up a wide variety of novel applications. Examples include lightweight cameras for mobile phones and other consumer electronics, lab-on-chip technologies, endoscopes small enough to be swallowed or inserted arthroscopically, and surgically implanted fluorescence microscopes only a few centimeters thick that can be mounted onto the head of a mouse or other small animals. More recently, lenses themselves have been miniaturized and made extremely thin using diffractive optics, metamaterials, and foveated imaging systems.

Notwithstanding this tremendous progress, today’s miniature cameras and microscopes remain constrained by the fundamental trade-offs of lens-based imaging systems. In particular, the maximum field-of-view (FOV), resolution, and light collection efficiency are all determined by the size of the lens(es). To further miniaturize microscopes while maintaining high performance, we must step outside the lens-based imaging paradigm.

Computational imaging has emerged as a powerful framework for overcoming the limitations of physical optics and realizing compact imaging systems with superior performance capabilities. Generally speaking, computational imaging systems employ a computational algorithm that relaxes the constraints on the imaging hardware. For example, super-resolution microscopes such as PALM and STORM use computation to overcome the physical limitations of the imager’s lenses.

At the extreme, computation can be used to eliminate lenses altogether and break free of the traditional design constraints of physical optics. The main principle of “lensless imaging” is to design complex but invertible transfer functions between the
incident light field and the sensor measurements \([3, 2, 4, 23, 16, 64, 65, 34]\). The acquired sensor measurements no longer constitute an image in the conventional sense but rather data that can be coupled with an appropriate inverse algorithm to reconstruct a focused image of the scene. This redefinition of the imaging problem significantly expands the design space and enables compact yet high-performance imaging systems. For example, amplitude or phase masks placed on top of a bare conventional image sensor combined with computational image reconstruction algorithms enable inexpensive and compact photography \([2, 4, 23, 22]\).

Lensless computational imaging has also enabled extremely lightweight and compact microscopes. For example, pioneering work has demonstrated that a bare image sensor coupled with a spacer layer and a recovery algorithm is a powerful tool for shadow imaging \([54]\), holography \([66, 67]\), and fluorescence \([21, 68]\) with applications in point-of-care diagnostics and high-throughput screening. A major advantage of lensless microscopy is its ability to substantially increase the FOV. In a lens-based microscope, the FOV is inversely proportional to the magnification squared, whereas in a lens-free system the FOV is limited only by the area of the imaging sensor \([44]\).

Despite the recent advances in ultra-compact lensless microscopy, one key application area has remained infeasible: micron-resolution three-dimensional (3D) imaging of incoherent sources (e.g., fluorescence) over volumes spanning several cubic millimeters. This application area is particularly relevant for fluorescent imaging of biological samples both in vitro and in vivo, where the illumination sources and image sensors must often be on the same side of the imaging target. The major challenge in 3D imaging of such incoherent sources is that in the absence of lenses, the incoherent point spread function lacks the high-frequency spatial information necessary for high-quality image reconstruction. While coherent sources can exploit the high-contrast
interference fringes that carry high-frequency spatial information, the absence of interference fringes when imaging incoherent sources makes direct extensions of holographic methods infeasible without additional optical elements [69].

Using a novel computational algorithm and optimized amplitude mask, we demonstrate the first 3D lensless microscopy technique that does not rely on optical coherence. The result is the world’s thinnest fluorescence microscope (less than 1 mm thick) that is capable of micron resolution over a volume of several cubic millimeters (Fig. 3.1). The major innovation that enables the “FlatScope” is an optimized 2D array of apertures (or amplitude mask) that modulates the incoherent point spread function, which makes high-frequency spatial information recoverable. In addition, the amplitude mask for FlatScope is designed to reduce the complexity of the image reconstruction algorithm such that the image can be recovered from the sensor measurements in near real-time. When an arbitrary amplitude mask is placed atop an $N \times N$ pixel image sensor, the transfer function relating the unknown image to the sensor measurements is an $N^2 \times N^2$ matrix containing $O(N^4)$ entries (e.g., a 1 megapixel image sensor produces a matrix with $\sim 10^{12}$ elements). The massive size of this matrix leads to two major impracticalities: first, calibrating such a microscope would require the estimation of $O(N^4)$ parameters, and second, reconstructing an image would require a matrix inversion involving roughly $O(N^6)$ computational complexity.
Figure 3.1: Traditional microscope vs. FlatScope. (A) Traditional microscopes capture the scene through an objective and tube lens (≈20 – 460 mm) resulting in a quality image directly on the imaging sensor. (B) FlatScope captures the scene through an amplitude mask and spacer (≈0.2 mm) and computationally reconstructs the image. Scale bar, 100 µm (inset, 50 µm). (C) Comparison of form factor and resolution for traditional lensed research microscopes, GRIN lens microscope, and FlatScope. FlatScope achieves high-resolution imaging while maintaining a large ratio of field of view (FOV) relative to the cross-sectional area of the device. (D) FlatScope prototype (shown without absorptive filter). Scale bars, 100 µm.
3.2 Imaging model

3.2.1 Amplitude mask design

Previous works [4, 34] have shown that using a separable amplitude mask drastically reduces the computational complexity. Motivated by the same, we use a similar design of separable mask pattern.

Two-dimensional (2D) separable mask patterns are constructed via the outer product of two 1D sequences, \( m_1 \) and \( m_2 \) of length \( N \), where each entry in the sequences is either \(-1\) or \(+1\). In the resulting 2D matrix, \(-1\) entries are assigned to closed apertures of the mask and \(+1\) entries are assigned to open apertures of the mask [4, 16]. Since closed apertures block light, they are instead assigned value 0 in the 2D matrix. The 2D mask pattern can be then written as the following matrix of dimensions \( N \times N \):

\[
M = \frac{1 + m_1 m_2^T}{2}
\]  

(3.1)

Specifically, we use a Modified Uniformly Redundant Array (MURA) [70] as our mask design. Modified Uniformly Redundant Array (MURA) has many attractive properties that have motivated the pattern’s use in astronomy, fast neutron, and gamma-ray imaging. Here we show for the first time that MURA is a separable pattern.

Let \( M \) represent the two-dimensional MURA pattern. Given a prime number \( p \) of the form \( p = 4m + 1 \), MURA is of size \( p \times p \). Let \( i = 0, \ldots, p - 1 \) index the rows and \( j = 0, \ldots, p - 1 \) index the columns of \( M \). Then the MURA pattern can be written
as (27):

$$M_{ij} = \begin{cases} 
0, & \text{if } i = 1 \\
1, & \text{if } j = 0, i \neq 0 \\
1, & \text{if } C_i C_j = +1 \\
0, & \text{if } C_i C_j = -1 
\end{cases}$$ \hspace{1cm} (3.2)

where,

$$C_i = \begin{cases} 
0, & \text{if } q \text{ is a quadratic residue modulo } p \\
-1, & \text{otherwise.} 
\end{cases}$$ \hspace{1cm} (3.3)

An integer $q$ is called a quadratic residue modulo $p$ if there exists an integer $x$ such that: $x^2 \mod p = q \mod p$.

To make MURA pattern $M$ separable, we drop the first row ($i = 0$) and the first column ($j = 0$) and redefine $M$ of size $(p - 1) \times (p - 1)$ as:

$$M_{ij} = \begin{cases} 
1, & \text{if } C_i C_j = +1 \\
0, & \text{if } C_i C_j = -1. 
\end{cases}$$ \hspace{1cm} (3.4)

If we define $C_I = \{C_i\}_{(i=1)}^p$ and $C_J = \{C_i\}_{(i=1)}^p$, then $M$ can be written in the separable form of Equation [3.1] as follows:

$$M = \frac{1 + C_I C_J^T}{2}$$ \hspace{1cm} (3.5)

3.2.2 Texas Two-Step model

Image formation for FlatScope can be described as follows. Let the scene, mask, and sensor be at the distances shown in Figure 3.2A, where $d_1$ is the distance between scene and mask, and $d_2$ is the distance between mask and sensor. Let the discretized scene be written as $X$, discretized mask as $M$ and discretized sensor measurement
as $Y$. For convenience, we will consider these quantities as matrices and index them by rows and columns. For example, $X(u, v)$ is the $u^{th}$ row and $v^{th}$ column of $X$. For convenience, we will assume that $X, M, Y$ are of the same dimension $N \times N$. In the following paragraphs, we will derive the image formation model for a scene at a single depth and later extend to multiple depths.

![Figure 3.2](image.png)

Figure 3.2: T2S derivation. (A) Illustration of scene, mask and sensor positions. (B) Sensor pixel’s 2D acceptance profile $C(u, v)$, which can be written as an outer product of the pixel’s response along rows ($c_1(u)$) and the pixel’s response along columns ($c_2(v)$).

When the scene element $X(u, v)$ is active (or illuminated), the sensor records a magnified version of the mask centered around $M(u, v)$ and scaled by the scene intensity $X(u, v)$. The magnification $\alpha = \frac{d_1 + d_2}{d_1}$ can be calculated using similar triangles as shown in Figure 3.2A. An additional term that needs to be considered is the sensor pixel’s response to light rays at different angles. Figure 3.2B shows the pixel’s response profile. Attenuation due to the pixel’s response causes the formation of a local pattern on the sensor when a point source, such as a fluorescent bead, is imaged by FlatScope (Figure 3.3A,B). We will denote the pixel’s response by $C$. The sensor
measurement at \((u', v')\) can then be written as:

\[
Y(u', v') = \sum_u \sum_v X(u, v)M\left(u + \frac{u' - u}{\alpha}, v + v' - v\right)C(u' - u, v' - v) \tag{3.6}
\]

The mask design procedure above enables us to write the following:

\[
M\left(u + \frac{u' - u}{\alpha}, v + v' - v\right) = \frac{1}{2} + \frac{m_1(u + \frac{u' - u}{\alpha})m_2(v + \frac{v' - v}{\alpha})}{2} \tag{3.7}
\]

Since the pixels are rectangular, each pixel’s response is separable as well:

\[
C(u' - u, v' - v) = c_1(u' - u)c_2(v' - v) \tag{3.8}
\]

where \(c_1\) is the pixel’s response along rows and \(c_2\) is the pixel’s response along columns (as show in Fig 3.2B). Substituting the above two equations in Equation 3.6 yields

\[
Y(u', v') = \sum_u \sum_v \left[c_1(u' - u)X(u, v)c_2(v' - v) + \frac{1}{2}c_1(u' - u)m_1\left(u + \frac{u' - u}{\alpha}\right)X(u, v)m_2\left(v + \frac{v' - v}{\alpha}\right)c_2(v' - v)\right]
\]

\[
= \frac{1}{2} \sum_u c_1(u' - u)\left[\sum_v X(u, v)c_2(v' - v)\right] + \frac{1}{2} \sum_u c_1(u' - u)m_1\left(u + \frac{u' - u}{\alpha}\right)\left[\sum_v X(u, v)m_2\left(v + \frac{v' - v}{\alpha}\right)c_2(v' - v)\right] \tag{3.9}
\]

Since the summation along the rows and summation along the columns have separated in the above equation, we can rewrite it concisely as

\[
Y = P_o X Q_o^T + P_c X Q_c^T \tag{3.10}
\]

where \(P_o\) and \(Q_o\) implement the effects of \(c_1\) and \(c_2\), respectively, and \(P_c\) and \(Q_c\) implement the effects of \(m_1\) and \(m_2\), respectively. Equation 3.10 shows that the sensor measurements can be written as a superposition of two separable functions. We call this superposition the **Texas Two-Step model (T2S)**.
Note that \( P_o \) and \( Q_o \) are not functions of the mask and, in fact, the first term in Equation 3.10 models the effect when an open (or no) mask is placed in front of the sensor. On the other hand, \( P_c \) and \( Q_c \) are functions of the mask pattern (i.e., \( m_1, m_2 \)) and hence the second separable term in Equation 3.10 models the effect due to the coding of the mask. The subscripts \( o \) and \( c \) refer to “open” and “coding”, respectively. This interpretation is visually shown in Figure 3.3. Also note that, \( P_o \) and \( Q_o \) have positive entries (since the sensor pixel’s response is always positive), while \( P_c \) and \( Q_c \) have both positive and negative entries.

The 1D sequences \( m_1 \) and \( m_2 \) are chosen to have half of their entries equal to \(-1/+1\). Therefore, \( m_1 \) and \( m_2 \) are orthogonal to a sequence with all +1 entries. Consequently, corresponding columns of \( P_o \) (no mask component) and \( P_c \) (coding component) are orthogonal. The same is true for corresponding columns of \( Q_o \) and \( Q_c \). We exploit this fact to expedite the calibration process.

### 3.2.3 Model complexity analysis

There are two aspects to consider when discussing computational tractability: memory requirements and runtime. Either large memory requirement or long run time is undesirable in a computational imaging system, particularly if real time processing is desired. In the following discussion, we will show that a separable mask that conforms to the T2S model requires far less memory and run time as compared to an arbitrary mask.

First, we consider the advantages of the T2S model in terms of the total number of parameters in the linear mapping from the scene to the measurements. This both reduces the effort for system calibration and speeds up the image reconstruction run time. To be general, consider a scene \( X \) and sensor measurement \( Y \) as matrices of size
Texas Two-Step Model: \[ Y = P_o X d Q_o d^T + P_c X d Q_c d^T \]

Figure 3.3: *Texas Two-Step model.* (A) Illustration of the FlatScope model using a single fluorescent bead as the scene. (B) Fluorescent point source at depth \( d \), represented as input image \( X_d \). (C) FlatScope measurement \( Y \). FlatScope measurement can be decomposed as a superposition of two patterns, (D) pattern when there is no mask in place (open) and (E) pattern due to the coding of the mask. Each of the patterns are separable along x- and y-directions and can be written as (F and G) two separable transfer functions. The FlatScope model, which we call as Texas Two-Step model (T2S), is the superposition of the two separable transfer functions.
For an arbitrary amplitude mask, the following generalized linear model holds:

\[ y = \Phi x \]  

(3.11)

where, \( x \) and \( y \) are the vectorized versions of \( X \) and \( Y \), respectively, formed by concatenating the columns of each matrix into a single, long vector. Both \( y \) and \( x \) are of length \( MN \), while \( \Phi \) is \( MN \times MN \). Therefore, \( \Phi \) contains \( M^2N^2 \) or \( O(N^4) \) elements if \( M \cong N \).

In contrast, a separable mask yields to the T2S model

\[ Y = P_o X Q_o^T + P_c X Q_c^T \]  

(3.12)

where, \( P_o \) and \( P_c \) are each of size \( M \times M \) and \( Q_o \) and \( Q_c \) are each of size \( N \times N \). This reduces the total number of elements in the mapping from the scene to measurements to \( 2(M^2 + N^2) \) or \( O(N^2) \) if \( M \cong N \).

Second, we consider the advantages of the T2S model in terms of the amount of memory required to store the parameters in the linear mapping from the scene to the measurements. For the FlatScope prototype presented here, \( M = 1000 \) and \( N = 1300 \). Hence, for an arbitrary (i.e., non-separable) mask, the matrix \( \Phi \) will contain 1.7 trillion elements. To be robust to quantization noise, we represent each element as a 32-bit floating point datatype. Therefore, storing the matrix \( \Phi \) would require 6 TB of memory, which is beyond the capabilities of both commercially available desktop computers and memory optimized cloud computing services (e.g., Amazon EC2 X1 currently offers memory of 2TB). This necessitates breaking \( \Phi \) into smaller chunks for computation, which drastically increases the data communication overhead between storage and RAM and supersedes any hardware speedups, yielding intractable run times of several weeks to months. Needless to say, multi-depth reconstruction that
requires multiple $\Phi$ matrices becomes practically impossible.

By using a separable mask, the total memory requirement for the matrices \{ $P_o, Q_o, P_c, Q_c$ \} in the T2S model is just 21 MB, a savings of five orders of magnitude over a non-separable approach. Due to low memory usage and reduced size of matrices, the gradient steps required for the iterative image reconstruction algorithms can be computed repeatedly with low computational cost. Additionally, we can use a parallel implementation on a GPU, which usually has less memory than a CPU. With a sub-optimal Nvidia Tesla GK210 GPU implementation using MATLAB, we achieved a single depth reconstruction in under 10 s and a 41-layer 3D reconstruction in under 15 min. An optimized GPU implementation will provide an even larger speedup.

3.2.4 Model calibration

FlatScope calibration relies on an observation that if the scene is separable (rank-1), then the FlatScope measurement is rank-2. For example, if the scene has only the $i^{th}$ row active (or illuminated), then the scene can be written as $X_i = e_i 1^T$, where $e_i$ is a sequence of zeros with only $i^{th}$ element to be 1 and $1^T$ is a sequence of all 1s. Then the FlatScope measurement can be written as:

$$ Y = (P_o e_i) (Q_o 1)^T + (P_c e_i) (Q_c 1)^T = p_{oi} q_o^T + p_{ci} q_c^T $$ (3.13)

Here, $p_{oi}$ and $p_{ci}$ are the $i^{th}$ columns of $P_o$ and $P_c$, respectively, and $q_o$ and $q_c$ are the sums of columns of $Q_o$ and $Q_c$, respectively. As noted in the Section 3.2.2, $p_{oi}$ and $p_{ci}$ are orthogonal and can be computed (up to a scaling factor) via the Singular Value Decomposition (SVD) of $Y_i$ truncated to the two largest singular values. Since the sensor measurements are always positive, the truncated SVD of $Y_i$ yields one
vector with all positive entries and another vector with both positive and negative entries. The positive vector is assigned to $p_{oi}$, and the other vector is assigned to $p_{ci}$. By scanning the rows of the scene, we can compute all the entries in $P_o$ and $P_c$.

Similarly, the columns of $Q_o$ and $Q_c$ can be calibrated by scanning along the columns of the scene. Scanning the rows and columns of the scene is physically done by translating a line slit, as shown in Figure 3.4. The transfer functions $\{P_o, Q_o, P_c, Q_c\}$ are dependent on the distance of the scene ($d_1$) and can be calibrated for each depth by first translating the line slit to the required depth and then scanning the FOV.

Figure 3.4: *Calibration setup.* (A) Side view of the calibration setup (not to scale) showing the LED array, wide angle diffuser ($\sim$10 cm above the LEDs), target with line slit ($\sim$1 cm above the diffuser) and FlatScope (filter, mask, spacer & sensor). (B) Calibration of rows by translating the horizontal slit target along the y-axis. (C) Calibration of columns by translating the vertical slit target along the x-axis.
3.3 Reconstruction algorithms

To be robust to various sources of noise, we formulate the reconstruction problem as a regularized least squares minimization.

$$\hat{X}_d = \arg\min_{X_d} \|P_{od}X_dQ_{od}^T + P_{cd}X_dQ_{cd}^T - Y\|_2^2 + \lambda_r \mathcal{R}(X_d)$$  \hspace{1cm} (3.14)

where $\mathcal{R}(\cdot)$ is the regularization chosen based on the scene.

For extended scenes like the USAF resolution target (Section ), we use Tikhonov regularization. For a given depth $d$ and calibrated matrices $P_{od}, Q_{od}, P_{cd}, Q_{cd}$, we estimate the scene $\hat{X}_d$ by solving a Tikhonov regularized least squares problem:

$$\hat{X}_d = \arg\min_{X_d} \|P_{od}X_dQ_{od}^T + P_{cd}X_dQ_{cd}^T - Y\|_2^2 + \lambda_2 \|X_d\|_2^2$$  \hspace{1cm} (3.15)

For sparse scenes (Section ), we solve the Lasso problem:

$$\hat{X}_d = \arg\min_{X_d} \|P_{od}X_dQ_{od}^T + P_{cd}X_dQ_{cd}^T - Y\|_2^2 + \lambda_2 \|X_d\|_1$$  \hspace{1cm} (3.16)

For the fluorescent samples (Sections ), we solve the 3D reconstruction problem as a Lasso problem:

$$\hat{X}_D = \arg\min_{X_D} \left\| \sum_{d=1}^{D} P_{od}X_dQ_{od}^T + P_{cd}X_dQ_{cd}^T - Y \right\|_2^2 + \lambda_1 \|X_D\|_1$$  \hspace{1cm} (3.17)

Iterative techniques and rudimentary GPU implementations were developed in Matlab to solve all of the above optimization problems. Equation 3.15 was solved using Nesterov’s gradient method [71], while Equations 3.16 and 3.17 were solved using FISTA (38) [72]. As expected, the longest running time was taken by the 3D deconvolution problem, with the solution converging in under 15 minutes.

3.3.1 Gradient direction for the T2S model

Since the T2S model is unlike a generic linear model, we show the gradient computation required for the afore-mentioned reconstruction algorithms.
For single depth reconstruction, let us define the forward operator $A_d(\cdot)$ and a transpose operator $A_{d}^T(\cdot)$ as follows:

$$A_d(X_d) = P_{od}X_dQ_{od}^T + P_{cd}X_dQ_{cd}^T$$  \hspace{1cm} (3.18)

$$A_d^T(Y) = P_{od}^T Y Q_{od} + P_{cd}^T Y Q_{cd}$$  \hspace{1cm} (3.19)

Then the gradient direction for optimization Equation 3.15 is:

$$\nabla = -2A_d^T(Y) + 2A_d^T(A_d(X_d)) + 2\lambda_2 X_d$$  \hspace{1cm} (3.20)

The gradient direction for optimization Equation 3.16 is:

$$\nabla = -2A_d^T(Y) + 2A_d^T(A_d(X_d))$$  \hspace{1cm} (3.21)

For 3D reconstruction, let us define the forward operator $A_D(\cdot)$ and a transpose operator $A_D^T(\cdot)$ as follows:

$$A_D(X_D) = \sum_{d=1}^{D} \left( P_{od}X_dQ_{od}^T + P_{cd}X_dQ_{cd}^T \right)$$  \hspace{1cm} (3.22)

$$A_D^T(Y) = \sum_{d=1}^{D} \left( P_{od}^T Y Q_{od} + P_{cd}^T Y Q_{cd} \right)$$  \hspace{1cm} (3.23)

Then the gradient direction for optimization Equation 3.17 is:

$$\nabla = -2A_D^T(Y) + 2A_D^T(A_D(X_D))$$  \hspace{1cm} (3.24)

3.4 Prototype

3.4.1 Imaging sensor and mask

To evaluate FlatScope’s performance, we constructed several prototypes. We describe the performance of a particular prototype based on a Sony IMX219 sensor with 2×2 pixel binning (using only the green pixels in a Bayer sensor array). We used a region
of 1300×1000 binned pixels to create an effective 1.3 megapixel sensor with 2.24 µm pixels. The amplitude mask is a 2D Modified Uniformly Redundant Arrays (MURA) \[70\] designed with prime number 3329, where the mask’s smallest feature size is 3 µm.

### 3.4.2 Fabrication

A 100 nm thin film of chromium was deposited onto a 170 µm thick fused silica glass wafer and then photolithographically patterned with a Shipley S1805. The chromium was then etched, leaving the MURA pattern with a minimum feature size of 3 µm. The wafer was diced to slightly larger than the active area of the imaging sensor. The imaging sensor Sony IMX219 provides direct access to the surface of the bare sensor for on-chip fabrication. The diced amplitude mask was aligned rotationally to the pixels of the imaging sensor under a microscope to enforce the separability under the T2S model and then epoxied to the sensor with Norland Optical Adhesive #72 using a flip chip die bonder. To filter blue excitation light in fluorescence experiments, we used an absorptive filter (Kodak Wratten #12) cut to the size of the mask and attached using epoxy with a flip-chip die bonder in the same manner. The device was finally conformally coated with < 1 µm of parylene for insulation. An overview of the fabrication process is shown in Figure 3.5.

### 3.4.3 Calibration

To calibrate FlatScope, we used 5 µm wide line slits fabricated in a 100 nm film of chromium on glass wafer and a LED array (green 5050 SMD) located ∼10 cm below the line slit (Figure 3.4A). To ensure that the light passing through the calibration slit was representative of a group of mutually incoherent point sources \[29\], we placed a
Figure 3.5: Fabrication of FlatScope. Steps for fabricating the amplitude mask and spacer, aligning and fixing to the imaging sensor, adding the absorptive filter and insulating/protecting the device.
wide-angle diffuser (Luminit 80°) between the target and light source as shown in the figure. While FlatScope remained static, the calibration slit, diffuser and LED array were translated with linear stages/stepper motors (Thorlabs LNR25ZFS/KST101) separately along the x-axis and y-axis (Figure 3.4B,C). The horizontal and vertical slits were translated over the FOV of the FlatScope, determined by the acceptance profile of the pixels in the imaging sensor (Figure 3.2B). The full-width-half-max of our sensor’s pixel acceptance angle is approximately 40°. The translation step distance of 2.5 µm was repeated at different depth planes ranging from 160 µm to 1025 µm (Thorlabs Z825B/TDC001), while a translation step distance of 1 µm was used for a single depth of 150 µm. This calibration needs to be performed only once and the calibrated matrices \[ \{P_{od}, Q_{od}, P_{cd}, Q_{cd}\} \quad d=1,2,\ldots,D \] can be reused as long as the mask and sensor retain their relative positions, i.e., not deformed.

### 3.5 Experimental results

#### 3.5.1 Lateral resolution

To test the lateral resolution of the FlatScope prototype we used test patterns composed of closely spaced double-slit resolution targets in a chrome mask with varying line spacing. We found that the lateral resolution of our prototype is less than two microns. In Figure 3.6 we show the FlatScope results with a target containing a slit gap of approximately 1.6 µm compared to a confocal image of the same target (Figure 3.6A). The FlatScope images were captured at a distance of 150 µm (Figure 3.6B) and resolve the gap (Figure 3.6C). A comparable microscope would require an objective with an NA \( \sim 0.16 \), typically found in objectives with 4-5× magnification. Note that while a 4× objective on a traditional lens-based microscope with a similar active
area sensor would provide a field of view of only $0.41 \text{ mm}^2$, FlatScope dramatically increases this field of view by more than ten times to $6.52 \text{ mm}^2$. These results show that FlatScope can produce high-resolution images while maintaining an ultra-wide field of view. Additionally, our computational algorithm allows for the incorporation of image statistics, enabling us to resolve features smaller than the size of the binned pixel, the minimum aperture on the mask, and the calibration line thickness ($2.24 \mu\text{m}, 3 \mu\text{m}, \text{and} 5 \mu\text{m}$, respectively for this prototype).

Double slits were fabricated in a 100 nm film of chromium on a silica glass wafer. As with calibration, we used a LED array and wide-angle diffuser to illuminate the target. We captured ground truth images with a confocal microscope (Nikon Eclipse Ti/Nikon CFI Plan Apo 10× objective), measuring the width of the slits to be $4.3 \mu\text{m}$ with a gap of $1.6 \mu\text{m}$. FlatScope images were captured at a distance of $150 \mu\text{m}$ with a 230 ms exposure; 5 images were averaged to increase the signal-to-noise ratio (SNR).

![Figure 3.6](image.png)

**Figure 3.6:** Lateral Resolution tests with double slits (A) Double slit with a $1.6 \mu\text{m}$ gap imaged with a 10× objective. (B) Captured FlatScope image. (C) FlatScope reconstruction of the double slit with a $1.6 \mu\text{m}$ gap.
3.5.2 Computational depth scanning

Because depth is a free parameter in our reconstruction algorithm, FlatScope can reconstruct focused images at arbitrary distances from the sample surface. To demonstrate this dynamic focusing, we captured images of a 1951 USAF target (Thorlabs R1S1L1N) at distances ranging from 200 µm to ∼1 mm. We expect some lateral resolution degradation as distance increases but predict that the information captured through the mask (along with our reconstruction algorithm) will aid in maintaining high-resolution through this range, especially when compared to no mask. The images reconstructed at the closest distance of 200 µm resulted in the best resolution (Figure 3.7A), while at a distance of ∼1 mm, line pairs in group 5 were still resolvable (Figure 3.7C). These results confirm the capability of FlatScope to resolve images over a significant distance range while still maintaining high resolution.

As with calibration, we used a LED array and wide-angle diffuser to illuminate the 1951 USAF target. Distances of 200 µm, 525 µm and 1025 µm were captured by translating along the z-axis (Thorlabs Z825B/TDC001). Exposure times were 24 ms, 28 ms, and 28 ms, respectively, and 5 images were averaged at each depth to increase the SNR.

3.5.3 2D Fluorescence imaging

We show imaging of a simple 2D fluorescence sample in Figure. The sample was constructed by drop casting a 100 µL sample of 10 µm polystyrene microspheres (1.7 x 10^5 beads/mL, FluoSpheres yellow-green) onto a microscope slide and fixed with a standard coverslip (∼170 µm thickness). A single image was captured with FlatScope 50 µm from the coverslip and using a 30 ms exposure. The excitation light for the fluorescence was provided by a 470 nm LED (Thorlabs M470L3) with filter
3.5.4 3D volume reconstruction

In addition to the ability to reconstruct images at arbitrary distances, FlatScope is also able to reconstruct an entire 3D volume from a single image capture. To showcase this ability, we prepared a 3D sample by suspending 10 μm fluorescent beads in an agarose solution. The 3D sample was prepared with the fluorescent beads (3.6 x 10^4 beads/mL) in a 1% agarose solution. A 100 μL portion of the mixture was placed onto a well slide and fixed with a standard coverslip (∼170 μm thickness).
Figure 3.8: 2D Fluorescence imaging of 10 µm fluorescent beads. (A) FlatScope reconstruction. (B) Reference image using an epifluorescence microscope used as ground truth. Scale bar, 100 µm (inset, 50 µm).

We captured a single image using FlatScope 50 µm from the coverslip and using a 30 ms exposure. The excitation light for the fluorescence was provided by a 470 nm LED (Thorlabs M470L3) with filter (BrightLine Basic 469/35) focused on the beads at an angle of ~60°. Reconstructed 3D volume from the single FlatScope image is shown in Figure 3.9A,B. As ground truth, we captured a 3D fluorescence volume using a scanning confocal microscope (Figure 3.9C). The ground truth images were captured with a depth range of 210 µm (and an area just > FOV of the FlatScope prototype) using a confocal microscope (Nikon Eclipse Ti/Nikon, CFI Plan Apo 10x objective). The confocal imaging required scanning and stitching to match the FOV of FlatScope; z-axis measurements were captured every 5 µm.

Comparing the FlatScope reconstruction to the confocal data over a depth range >200 µm, we not only have sufficient lateral resolution to resolve the beads, but we also achieve an axial resolution comparable to that of the confocal microscope.
Empirically, we find that the axial spread of the 10 µm beads is approximately 15 µm in the FlatScope reconstructions (Figure 3.9D).

Two important advantages of FlatScope’s ability to image complete 3D volumes from a single capture are data compression and speed. To obtain 3D volumes comparable to FlatScope, a confocal microscope must scan in the both the lateral and axial dimensions, capturing a series of images one pixel at a time. As a result, a large amount of data must be collected. For example, to image the fluorescent beads in Figure 3.9C, the confocal microscope had to overlay a total of 41 z-sections, with each z-section imaged at 4 megapixels, for a total of 164 recorded megavoxels. In contrast, FlatScope can achieve the same depth resolution with a single capture of 1.3 megapixels, which represents a 41× data compression (from depth alone). Moreover, the confocal data collection took more than 20 minutes, while FlatScope’s capture took 30 ms (a 40,000× speed-up). It must be noted that while a confocal microscope can achieve diffraction-limited lateral resolution, the lateral resolution of FlatScope is limited to about 2 µm in our prototype due to the pixel size on the image sensor we used. The FlatScope resolution can be improved as pixel sizes on image sensors become smaller.

3.5.5 3D volumetric video

FlatScope’s radically reduced capture time enables 3D volumetric video capture at the native frame rate of the image sensor. To demonstrate this capability, we recorded video of 10 µm beads (3.6 x 10^6 beads/mL) flowing through a 3D microfluidic device with two channels. FlatScope video was captured for 640×480 pixels (2×2 binned) at 18 frames-per-second at a distance of ~100 µm from the coverslip (~150 µm thickness) on which the microfluidic device was mounted. The microfluidic channels
Figure 3.9: 3D Volume Reconstruction of 10 µm fluorescent beads suspended in agarose.  
(A) FlatScope reconstruction is shown as a maximum intensity projection along the z-axis as well as a ZY slice (blue box) and an XZ slice (red box). (B) Estimated 3D positions of beads from the FlatScope reconstruction. (C) Reference data from a confocal microscope used as ground truth. (D) The depth profile of reconstructed beads compared to ground truth confocal images. Empirically, we can see that the axial spread of 10 µm beads is around 15 µm in FlatScope reconstruction. That is, FlatScope’s depth resolution is less than 15 µm. The three beads shown are at depths of 255 µm, 270 µm and 310 µm from the top surface (filter) of the FlatScope.
had approximate dimensions of 50 \( \mu m \times 40 \mu m \), with an axial separation of the channels of \( \sim 100 \mu m \). The excitation light for the fluorescence was provided by a 470 nm LED (Thorlabs M470L3) with filter (BrightLine Basic 469/35) focused on the beads at an angle of \( \sim 60^\circ \).

Figure 3.10 shows a time-lapse from a movie of a sub-section of the FlatScope 3D volume reconstruction and corresponding still-frames of the fluorescent beads as they travel through the two channels.

Figure 3.10: 3D Volumetric Video Reconstruction of moving 10 \( \mu m \) fluorescent beads. (A) Sub-section of FlatScope time-lapse reconstruction of 3D volume with 10 \( \mu m \) beads flowing in microfluidic channels (approximate location of channels drawn to highlight bead path and depth). Scale bar, 50 \( \mu m \) (FlatScope prototype graphic at top not to scale). (B) Captured images of frames 1, 3 and 5 (false colored to match time progression). (C and D) FlatScope reconstructions of frames 1, 3 and 5 at estimated depths of 265 \( \mu m \) and 355 \( \mu m \), respectively (dashed lines indicate approximate location of microfluidic channels). Beads and reconstruction false colored to match time progression. Scale bar, 50 \( \mu m \).
3.6 Discussion

We have demonstrated single-frame, wide-field 3D fluorescence microscopy with FlatScope over depths of hundreds of microns with the ability to refocus a single capture beyond 1 mm. Importantly, FlatScope achieves this unprecedented performance without a lens and with all optics and filters residing in a layer less than 300 $\mu$m thick on top of a bare imaging sensor. This results in a remarkably compact form-factor: our FlatScope prototype features lateral dimensions of 8 mm $\times$ 8 mm, thickness below 600 $\mu$m and weight of $\sim$0.2 grams. To our knowledge, no other compact lensless device is capable of 3D fluorescence imaging over hundreds of microns in depth. Though the focus of FlatScope in this work is fluorescence, the FlatScope concept is also applicable to bright-field, dark-field, and reflected-light microscopy.

The dynamic range of our FlatScope fluorescence images was degraded by the autofluorescence of the gel filter (see Methods), losing up to 64% of sensor’s dynamic range due to autofluorescence of the filter. Using an alternative thin-film absorptive filter (30, 31), potentially in conjunction with an omnidirectional reflector(32, 33), we expect to substantially increase the dynamic range of FlatScope while reducing its thickness to less than 200 $\mu$m. This increase in dynamic range will enable improved imaging of biological samples with the goal of in vivo 3D fluorescence imaging. We are also planning to integrate $\mu$LEDs around the imaging sensor to remove the need for an external excitation source. Fabrication advancements have led to $\mu$LEDs that are less than the thickness of our spacer layer which will enable us to maintain our overall compact form factor. Because the FOV of FlatScope is limited only by the sensor size, larger imaging sensors can easily be integrated for extreme wide-field imaging. Additionally, the small footprint of FlatScope raises the exciting prospect of arraying FlatScopes on flexible substrates for conformal imaging. Overall, FlatScope
opens up a new design space for 3D fluorescence microscopy where size, weight, and performance are no longer determined by the optical properties of physical lenses.
Appendices

3.A Comparison with FlatCam model

Given a separable mask, the T2S model applies at all scene distances. At large working distances, it can be simplified to the FlatCam (14, 16) imaging model, which features only one and not two separable terms. At large working distances, each scene element acts like a source of plane waves resulting in a global effect on the sensor as opposed to a localized effect. The global effect renders the open mask component ($P_oXQ^T_o$ term in Equation 3.10) of the T2S model as a scene-dependent constant. This constant can be subtracted using a heuristic method (like averaging), yielding a single separable model: $Y_{\text{FlatCam}} = P_cXQ^T_c$ (second term in Equation 3.10).

FlatCam’s single separable model, as proposed by Asif et al. (14), is an approximation to the T2S model and holds true only for large scene distances. The invariance of the T2S model to scene distance and the restricted capability of the FlatCam model are shown through simulation in Figure 3.B.1. Note that the simulated scene in the figure is fairly dense; for a sparser scene, the distance for a quality reconstruction using FlatCam would be far greater than the 2.5 mm shown in the figure.

3.B Simulation comparison

We also compare, by simulation, the performance of imaging incoherent light sources of FlatScope vs. two other lensless camera designs that share similar FOV and form-factor advantages over traditional lens-based microscopes (Fig 3.B.1). Sencan
et al. (28) have shown that using a bare image sensor without a mask is cost-effective for fluorescence lab-on-chip applications. To achieve high spatial resolution, sparsity constraints and close proximity to the bare sensor must be enforced so that the reconstruction (deconvolution) remains stable and tractable (29). The quality of images reconstructed from bare sensor images decays rapidly with increasing depth (Figure 3.B.1(A)). In contrast, the single separable model proposed by Asif et al. (14) breaks down as expected when the source distance approaches the device size, i.e., in the regime where the point spread function of a point source is localized to a portion of the sensor (see Supplementary Section 3.A). In this regime, our new T2S model is a necessity for high-resolution reconstructions (Figure 3.B.1(B,C)).
Figure 3.B.1: Simulation comparison of bare sensor, FlatScope, and FlatCam. Reconstructions are shown at depths 0.083 mm, 1.5 mm and 2.5 mm from respective systems. (A) Reconstruction of bare sensor decays rapidly with increasing depth. Reconstruction of FlatScope shows high resolution and stays fairly stable with increasing depth. (C) FlatCam is unable to reconstruct at smaller depths due to limitations of the model with a single separable term. On the other hand, the T2S model of FlatScope can handle all depths. At larger depths, FlatCam and FlatScope reconstructions are comparable.
3.C  Refractive index matching

The separability of the FlatScope model is based on light propagation through a homogeneous medium. A large change in refractive index across the target medium to the FlatScope interface (e.g., air to glass) results in a mapping of lines in the scene to curves at the sensor plane. Since separability requires the preservation of rectangular features, the curving effect weakens the T2S model. A small change in refractive index (e.g., water to glass) was observed to only minimally affect the model (Figure 3.C.1). For calibration and the experiments presented, a refractive index matching immersion oil (Cargille #50350) was used between the surface of the mask and the target.

Figure 3.C.1: Refractive index matching. Comparison of reconstructed images captured with different media between the surface of the mask and the target, (A) immersion oil, (B) H\textsubscript{2}O, and (C) air. Images were captured at a depth of 200 µm. Note that all images for the calibration process were captured with immersion oil between the surface of the mask and the target. Scale bar, 100 µm.
3.D Aberration removal before reconstruction

In practice, the FlatScope prototype could suffer from the following unwanted artifacts: dead or saturated pixels on the image sensor, dust trapped beneath the spacer, and air bubbles trapped in epoxy. We refer to these artifacts, generally, as “aberrations.” Aberrations do not fit into the T2S model since they are not separable and are invariant to the scene/sample. Hence, aberrations act as strong noise in localized regions and result in erroneous reconstruction around these regions. To correct aberrations, it is not sufficient to subtract them from the measurements; we also need to fill in the appropriate regions with correct values. To achieve the correction, we observe that our captured images are fairly low-dimensional (low rank when the captured image is considered a matrix), and aberrations occur as sparse outliers. Robust PCA (35, 36) is an effective algorithm to separate such sparse outliers from an inherently low-dimensional matrix. We use Robust PCA as a pre-reconstruction processing step to replace aberrations with aberration-corrected sensor values in the captured image. See Figure 3.D.1 for an experimental example.

3.E Filter autofluorescence compensation

The Kodak Wratten filter autofluorescence induces a significant DC shift in the captured images. Given the limited dynamic range of the sensor, this leads to contrast loss. We can remove the DC shift by subtracting the mean of the captured image, however, the filter does not autofluoresce uniformly invalidating an exact DC shift assumption. Since the brightness profile due to autofluorescence is of spatially low frequency, we subtract the two lowest frequency components as obtained by a Discrete Cosine Transform (DCT) decomposition of the image. Subtracting the lowest frequency com-
Figure 3.D.1: Aberration removal using Robust PCA. (A) Captured FlatScope image before aberration removal. (B) Processed capture without aberrations. (C) Aberrations removed by using Robust PCA. (D) Reconstructed image prior to Robust PCA shows large artifacts. (E) Processed reconstruction with aberration removal results in substantially reduced artifacts. (F) Ground truth image. Scale bars, 50 µm.
ponents eliminates almost all of autofluorescence (Fig. 3.E.1A,B). We note that up to 64\% of the dynamic range of the sensor is lost to autofluorescence, which limits the performance of FlatScope by reducing the signal strength. Despite this limitation, we are able to show high-quality image reconstructions (Fig. 3.E.1C,D)). Shifting to a thinner and less autofluorescent filter will improve the performance of FlatScope.

Figure 3.E.1: Removing autofluorescence using DCT. (A) Captured FlatScope image before DCT removal. (B) Processed capture image. (C) Reconstructed image prior to DCT removal shows excessive noise. (D) Processed reconstruction with DCT removal results in substantially reduced noise. Scale bars, 100 \(\mu\)m.
3.F  Diffraction effects on T2S model

The derivation of the T2S model in section 3.2.2 is based on geometric optics; but at the scale of our mask features, diffraction effects from wave optics can have a significant impact. To quantify the errors caused when not incorporating wave optics, we compare the raw point spread function (PSF) of our prototype with the T2S approximation of the PSF. The comparison is shown in Figure 3.F.1. The raw PSF is captured by illuminating a 5 \( \mu \)m precision pinhole (Edmunds Optics, #38-537) with a wide-angle diffuser. The pinhole acts like a point source when illuminated with a diffuser and is placed at a depth 250 \( \mu \)m from the FlatScope prototype. We see from Figure 3.F.1 that there is a small difference between the raw captured PSF and the T2S approximation, with an average error of 12.6%. In our experiments, we use the raw captured images for reconstruction. Notwithstanding the geometrical approximation, we can reliably reconstruct scenes (as shown in Figs. 3.6, 3.7, 3.8, 3.9 and 3.10) using the T2S model.

In the next chapter, we overcome mismatch due to diffraction by incorporating diffraction effects into the mas design.

3.G  Impact of sensor saturation

In the absence of a focusing lens, the light from any point in a fluorescent sample is spread across multiple pixels. Hence, for a given exposure duration (those common in fluorescence microscopy), a lensed system might saturate, but FlatScope would likely not. If saturation occurs in a specific region, the artifacts of reconstruction are localized around that region and don’t affect farther locations. In practice, we want to avoid any nonlinear effects like saturation since our model and reconstructions are
Figure 3.F.1: T2S model error from diffraction. (A) Raw capture by FlatScope prototype of a point source placed at a depth of 250 µm. The raw capture is cropped to the region of interest. (B) T2S model approximation of the capture. (C) Mask region illuminated by the point source. The location of the point source was chosen such that the illuminated mask region has a diversity in aperture sizes (white represents open apertures; black represents closed apertures). (D) Absolute error difference between raw capture and T2S model approximation. Average error is 12.64%. Scale bar, 40 µm.
based on a linear sensor response. In the future, we may borrow ideas from high
dynamic range (HDR) photography to combine multi-exposure captures to ensure
quality reconstruction for very high contrast samples.
Chapter 4

Light effiecent high-resolution lensless imaging

4.1 Introduction

For decades lenses have been studied and designed to achieve high resolution sharp images on the sensor. But, using lens in a camera system comes with a few physical constraints. A lens needs to be placed at least a focal length away from an imaging sensor, which places a minimum camera width constraint. Lenses aren’t flat and hence cannot be integrated with flat surfaces. Lensless cameras can be very thin, flat and could be arrayed without being chunky. Giving up lenses means giving up capturing recognizable images on the sensor. With lack of lenses, sensor captures are muddled and lack any resemblance to the scene being imaged \[16\]. Fortunately, advances in computation algorithms have made it possible to invert such muddled sensor captures to recognizable high quality images.

To go completely optics free, researchers have shown imaging with a bare sensor \[23\]. They exploit shadows cast by defects on the sensor’s cover glass, and the anisotropy of pixel response to achieve some level of light modulation. Due to limited control on the light modulation, the achievable reconstruction quality is limited to very low resolutions. Another way to achieve bare sensor imaging is to have an active illumination that can produce coding or light modulation from object shadows \[54\] or from interference fringes \[67\]. But adding an active illumination makes the imaging system bulkier, beating the purpose of size reduction.
To passively achieve a higher control of light modulation, researchers have built lensless imaging systems by placing carefully designed “masks” in front of an imaging sensor [16]. The most popular choice for masks are binary amplitude masks [4, 34, 5]. The light modulation is achieved through shadow of the amplitude mask which behaves as the point-spread-function (PSF) of the system. To achieve a particular binary PSF, the amplitude mask is designed with the same pattern as the PSF. A huge issue with amplitude mask is that they block a significant amount of light which can lead to very low SNR in low light scenarios.

Another issue with amplitude mask is that the space of achievable PSFs are limited by diffraction. This limits the minimum feature size achievable in a PSF. Smaller features are desired in PSF to achieve higher resolution [4, 5]. Also, if the PSF is designed to have some desirable properties like being separable [4], generating moire fringes [5], and having flat Fourier spectrum, diffraction effects will suppress these properties. As a rule of thumb, the Fresnel number $N_F$ associated with the amplitude mask can help in determining whether the PSF will be close to the pattern of the mask or different [29]. If the Fresnel number is much greater than 1, then geometrical properties are valid and the shadow PSF mimics the mask pattern. Hence, for mask to sensor distance of $1200 \, \mu m$, the smallest aperture on the mask needs to be greater than $60 \, \mu m$ to achieve a reliable shadow pattern. This means that the PSF cannot be designed to have a feature size smaller than $60 \, \mu m$. The difference in amplitude mask pattern and the generated PSF can be seen in Fig. 2. We would like to point out that although PSF due to amplitude mask can reach smaller feature sizes from diffraction, these features don’t mimic the mask pattern, hence undesirable. We could, however, design amplitude masks to achieve desired PSF by incorporating diffraction [73], but the loss of light makes this approach unappealing for imaging.
Figure 1: An outline of mask design and system architecture for our miniature diffractive camera. We first obtain a target point spread function (PSF) through PSF engineering. Then, using our iterative phase mask design, we optimize for a phase map that can reliably produce the target PSF.

In this chapter, we propose the design and use of phase mask that reliably produces the desired target PSF. A phase mask modulates the phase of light and doesn’t block any light, thereby is the most light efficient mask choice for a lensless camera. Additionally, the proposed design of phase mask incorporates diffraction effects or, specifically, wave optics to achieve a wider space of PSF designs that are beneficial in improving the performance of a lensless imager.

Our design of lensless cameras consists of two parts: (1) PSF engineering and (2) phase mask optimization for the chosen PSF. We propose a high-performance contour PSF and validate our design through experiments. An illustration of our design process is shown in Fig. 1.

Our contributions are:

1. An optimization algorithm to generate phase masks (aka diffractive optical elements) for any device thickness and given point spread function.

2. An optimized contour-based point spread function that is efficient for imaging application and reaches the theoretical resolution limit for any given thickness.
4.2 Diffractive masks

Diffractive masks used for lensless cameras can be broadly categorized into amplitude masks and phase masks. Amplitude masks were used by [4, 17, 34, 19] and phase masks were used by [74, 75, 18].

4.2.1 Amplitude Masks

An amplitude mask modulates the amplitude of incident light by either passing, blocking or attenuating. For ease of fabrication, a binary amplitude is commonly used. The light modulation is achieved through the shadow of the amplitude mask, which behaves as the PSF of the system.

A concerning issue in using amplitude mask is the light throughput. Since the mask modulates light by passing and blocking light, a significant amount of photons are lost leading to low signal-to-noise-ratio (SNR) sensor capture. Low SNR is undesirable for low light scenarios and photon-limited imaging like fluorescence or bioluminescence imaging. Additionally, decoding the lensless sensor capture tends to amplify noise leading to poor reconstruction. Amplitude masks also suffer from diffraction effects which limit the range of achievable PSFs.

Design

Among all masks, designing an amplitude mask based on a desirable PSF is the most straightforward. The pattern of a binary amplitude mask is merely the PSF itself, where the bright regions of the PSF correspond to the open areas of the mask and the dark regions correspond to the blocking areas [16].
However, the range of PSFs achievable using the above mentioned, ray optics based, amplitude mask design is limited due to diffraction. As a rule of thumb, the Fresnel number $N_F$ [29] associated with the amplitude mask can help in determining whether the PSF will be close to the pattern of the mask or different. If the Fresnel number is much greater than 1, then geometrical properties are valid and the shadow PSF mimics the mask pattern. When the Fresnel number falls below 1, the cast PSF will deviate from the mask pattern. The difference in amplitude mask pattern and the generated PSF can be seen in Fig. 2.

![Figure 2: Target and achieved PSF using (a) binary amplitude mask and (b) phase mask. The PSF from amplitude mask is heavily degraded by diffraction effects.](image)

4.2.2 Phase Masks

A phase mask modulates the phase of incident light by the principles of wave optics [76]. Phase masks allow most of the light to pass through providing high SNR. Hence, they are desirable for low light scenarios and photon-limited imaging.

We propose to use phase mask for our lensless camera and present a mask design algorithm to achieve desirable PSFs.
Design

Odd-symmetry phase gratings were proposed by [7] to achieve robust nulls in the PSF. The wavelength and depth-robust nulls are produced along the axis around which phase gratings have an odd-symmetry. However, this design doesn’t guarantee intensity distributions in the non-null regions of the PSF.

Use of diffuser as a phase mask for lensless imaging was proposed by [18]. To achieve the best performance from a diffuser, it has to be placed at a distance from the sensor such that the pattern projected is at the highest contrast. Since the phase profile of a diffuser is inherently statistical, the optimal distance varies from one to another.

Phase mask design using phase retrieval algorithms was proposed by [77] and implemented using phase spatial light modulator in [74]. We follow a similar approach in designing our phase mask for a desirable PSF.

4.3 Diffractive Camera Model

4.3.1 Imaging Architecture

Our lensless camera has a fabricated diffractive element called phase mask placed at a distance $d$ from an imaging sensor (Figure 3(a)). A phase mask modulates the phase of incident light and produces a pattern at the sensor through constructive and destructive interference. In the following sections, we’ll describe how the phase mask produces interference pattern and the consequent diffractive imaging model.
4.3.2 Fresnel Propagation

When a phase mask, with phase profile $\phi(\xi, \eta)$, is illuminated with a coherent collimated light, the intensity pattern $p(x, y)$ captured by the imaging sensor placed at distance $d$ is given by magnitude square of Fresnel propagation \cite{29}:

$$p(x, y) = |F_d(e^{j\phi(\xi, \eta)})|^2$$

$$= \left| \frac{1}{j\lambda d} \int \int e^{j\phi(\xi, \eta)} \exp \left[ j \frac{\pi}{\lambda d} \left( (x - \xi)^2 + (y - \eta)^2 \right) \right] d\xi d\eta \right|^2,$$  \hspace{1cm} (4.1)

where $F_d(\cdot)$ denotes fresnel propagation by distance $d$ and $\lambda$ is the wavelength of light. For simplicity, let’s consider a one-dimensional (1D) phase mask with phase map $\phi(\xi)$ and drop the scaling term. Then, the pattern produced from collimated light parallel to optical axis is given by

$$p(x) = \left| \int e^{j\phi(\xi)} \exp \left[ j \frac{\pi}{\lambda d} (x - \xi)^2 \right] d\xi \right|^2$$

$$= \left| \int e^{j\phi(\xi)} \exp \left[ j \frac{\pi}{\lambda d} \left( \xi^2 - 2x\xi \right) \right] d\xi \right|^2,$$  \hspace{1cm} (4.2)

where the quadratic term was expanded and a constant phase term was removed since we are considering only intensity.

4.3.3 Point Spread Function

The collimated light or planar waves can be said to be generated from an on-axis point source at a sufficiently large distance from the mask. Then, $p(x)$ (or $p(x, y)$ for 2D) can be called as the point-spread-function (PSF) of the system.

If the point source is off-axis, illuminating the phase mask at an angle $\theta$, it imparts
a linear phase $e^{j \frac{2\pi}{\lambda} \sin \theta}$ to Eq. 4.2 and the resultant intensity pattern is

$$I_\theta(x) = \left| \int e^{j \frac{2\pi}{\lambda} \sin \theta} e^{j \phi(\xi)} \exp \left[ j \frac{\pi}{\lambda d} (\xi^2 - 2x\xi) \right] d\xi \right|^2$$

$$= \left| \int e^{j \phi(\xi)} \exp \left[ j \frac{\pi}{\lambda d} (\xi^2 - 2(x - d \sin \theta)\xi) \right] d\xi \right|^2$$

$$= p(x - d \sin \theta).$$

Hence, an off-axis point source causes a lateral shift of the PSF. If the point source is at a distance $z_\infty$ and height $x_h$ then by paraxial approximation

$$\delta_{z_\infty}(x - x_h) \rightarrow p \left( x - \frac{d}{z_\infty}x_h \right), \quad (4.3)$$

where $\delta_z(x)$ denotes point source at distance $z$ from the mask. The shift property is illustrated in Figure 3(b) and can be stated as

Property 1

Shift invariance: A lateral shift of point source causes translation of PSF on the sensor plane.

The above property is also called “memory effect” [78, 79] and was recently used to perform non-invasive imaging through scattering media [80, 81, 82] and wavefront
sensing \[83\]. As we will see later, the shift invariance property helps us to write the imaging model as a convolutional model.

For a point source at a finite distance \(z\) from the mask, it imparts an additional quadratic phase \(e^{j \frac{\pi}{\lambda z} \xi^2}\) to Eq. 4.2 to give an intensity response as:

\[
I_z(x) = \left| \int e^{j \frac{\pi}{\lambda d} \xi^2} e^{j \phi(\xi)} \exp \left[ j \frac{\pi}{\lambda d} \left( \xi^2 - 2x\xi \right) \right] d\xi \right|^2
\]

\[
= \left| \int e^{j \phi(\xi)} \exp \left[ j \frac{\pi}{\lambda d} \left( \xi^2 - 2x \frac{x}{1 + d/z} \xi \right) \right] d\xi \right|^2
\]

\[
\approx \left| \int e^{j \phi(\xi)} \exp \left[ j \frac{\pi}{\lambda d} \left( \xi^2 - 2x \frac{x}{1 + d/z} \xi \right) \right] d\xi \right|^2
\]

\[
= p\left( \frac{x}{1 + d/z} \right).
\]

Here we assumed \(d \ll z\) which would be the case for our lensless cameras. Therefore, following the same notations as Eq. 4.3 we have

\[
\delta_z(x) \rightarrow p_z(x) = p\left( \frac{x}{1 + d/z} \right), \quad (4.4)
\]

which is a geometrical magnification of the PSF. The magnification property is illustrated in Figure 3(c) and can be stated as

Property 2

An axial shift of point source causes magnification of PSF on the sensor plane.

4.3.4 Convolutional model

A real world 2D scene \(i(x, y; z)\) at distance \(z\) can be assumed to be made up of incoherent point sources. Each point source will produce a shifted version of PSF \(p_z(x, y)\) and since the sources are incoherent to each other, the shifted PSFs will add linearly in intensity \[29\] at the sensor. By Property \[\text{II}\] of the PSF, we can write the imaging model as following convolution model:

\[
b(x, y) = p_z(x, y) * i(x, y; z) \quad (4.5)
\]
Here, \( b \) is the sensor’s capture and \( * \) denotes 2D convolution over \((x, y)\).

Different PSFs are optimal for different applications. Once a PSF is engineered for a particular application, we need a mechanism to realize this PSF with a diffractive mask. In the next section, we will discuss a optimization framework to realize the desired PSF with a phase mask.

### 4.4 Diffractive mask design

Given a target point spread function (PSF) and any system-related constraints such as constraints on the device thickness, the main goal of this step is to compute the optimal mask that can achieve the target point-spread functions within system constraints. We do this by first relating the point spread function and the phase mask analytically and then optimizing for the phase mask that maximally achieves that PSF.

From scalar diffraction theory, we know that given a phase mask profile \( \phi(\xi, \eta) \), the PSF \( p(x, y) \) generated on a CMOS sensor at distance \( d \) from the phase mask is given by Fresnel propagation \([29]\) (Eq. 1.1). Our desire is to design a phase mask that can produce the engineered PSF. Such problem of optimizing a phase profile that can produce an intensity pattern, such as the PSF, can be considered as a phase retrieval problem \([47]\). Here we adopt an approach similar to the Gerchberg-Saxton (GS) algorithm \([84]\) and motivated by \([77]\). The major difference compared to the GS algorithm is that our distance of wave propagation is in the near-field due to the absence of lens and the small camera thickness. Hence, the propagation of consideration here is not Fraunhofer diffraction \([29]\) (implemented with just a Fourier transform) but Fresnel propagation \([29]\) (has an additional quadratic phase). Nevertheless, we find that the near-field phase retrieval algorithm converges for a variety of PSFs as
we will see in the Results section.

The near-field phase retrieval algorithm for phase mask optimization can be described as follows. The algorithm uses an iterative approach, iterating between the fields at the mask plane and the sensor plane while simultaneously enforcing constraints at the two planes. The constraints are: the amplitude of the field at the mask plane is unity and intensity of field at the sensor plane is the target engineered PSF. The iterative approach is summarized in Supplementary Alg. 1 and Fig. 4. The phase mask optimization requires the following inputs: the target PSF, mask to sensor distance, the wavelength of light and the index of refraction of the mask substrate.

**Algorithm 1 Phase mask optimization**

**Input:** Target PSF $p(x, y)$, wavelength $\lambda$, mask to sensor distance $d$ and refractive index $n$ of mask substrate.

**Output:** Mask’s phase profile $\phi(\xi, \eta)$ and height map $h(\xi, \eta)$.

$M_p \leftarrow \sqrt{p}$

repeat

$M_\phi \leftarrow \mathcal{F}_{-d}(M_p)$ \{Back propagate from sensor to mask\}

$\phi \leftarrow \text{phase}(M_\phi)$

$M_\phi \leftarrow e^{i\phi}$ \{Constrain amplitude of mask to be unity\}

$M_p \leftarrow \mathcal{F}_d(M_\phi)$ \{Forward propagate from mask to sensor\}

$M_p \leftarrow \sqrt{p} \odot e^{i\text{phase}(M_p)}$ \{Constrain amplitude to be $\sqrt{\text{PSF}}$\}

until maximum iterations
Figure 4 : Phase mask optimization.
4.5 PSF Engineering

4.5.1 Previous lensless imaging PSFs

Various PSFs have been used for lensless imaging for their attractive properties. We describe them below.

Separable PSF

A separable PSF is constructed by an outer product of two 1-D vectors. Such construction simplifies the imaging model as convolution along the rows of the image followed by convolution along the columns. In matrix form, this operation can be written as a product of 2-D image with a few small 2-D matrices \[4, 34, 17\]. An example of separable PSF is shown in Fig. 5(a), constructed from outer product of two maximum length sequences \[85, 86, 4\].

Fresnel Zone Aperture

A Fresnel Zone Aperture (FZA) PSF is constructed like a Fresnel zone plate \[87\] and was used by \[5, 19\]. Multiplying the sensor capture with a virtual FZA results in overlapping moiré fringes. Fast reconstruction is done by applying a 2-D Fourier transform on the moiré fringes. An example of FZA PSF is shown in Fig. 5(b).

Spiral PSF

A spiral PSF was proposed by \[7\]. To cover a large sensor area, \[6\] proposed tessel-lating the spiral PSF. An example of the tessellated spiral PSF is shown in Fig. 5(c). A single unit of the tessellation is shown on the top left corner.
Diffuser Caustics

Using caustics from a diffuser as lensless PSF was proposed by [18]. Experimental capture of the diffuser PSF is shown in Fig. 7(a).

4.5.2 Performance and PSF

A lensless camera encodes an image onto the sensor through the PSF. We analyze the invertibility of various lensless camera PSFs by studying their frequency spectrum and noise performance. In addition, we propose a contour-based PSF that yields the best imaging performance.

Deconvolution involves inversion of the frequency spectrum, hence having large and almost flat magnitude spectrum is desirable in a PSF. Thus, performance of any PSF can be compared using their frequency spectrum.

We compare PSFs using the following metrics:

1. **Magnitude Spectrum** shown in Fig. 5(e). Magnitude spectrum remaining large and flat throughout all frequencies is a desirable property of a high performing PSF.

2. **Singular Values** of associated convolution matrix, shown in Fig. 5(f). Singular Values of a convolution matrix is the magnitude spectrum of the PSF ordered in descending order. Singular values remaining large and flat throughout is a desirable property of a high performing PSF.

3. **Mean Squared Error** normalized over noise variance, shown in Fig. 5(g). Mean Squared Error (MSE) performance of a PSF can be calculated as the sum of inverse of singular values [88]. Low values for MSE indicate higher signal to noise ratio of the reconstruction yielding to a high performing PSF.
Figure 5: Analysis of various lensless point-spread-functions (PSF). (a) Separable PSF created from outer product of two m-sequences as used by [4]. (b) Fresnel zone apertures (FZA) as used by [5]. (c) Tessellated spiral PSF as used by [6]. A single unit [7] of the tessellation is shown on the top left corner. Tessellation is needed to cover a larger sensor area. (d) Our proposed Perlin noise based PSF. (e) Magnitude spectrum of each PSF. The magnitude spectrum of Contour PSF remains large for entire frequency range indicating better invertibility characteristic. (f) Singular values of convolution matrix for each PSF. Singular value of Contour PSF remains fairly larger than for other PSFs indicating high performance. The large initial values for the spiral PSF is due to frequency harmonics caused by tessellation. (g) Mean Squared Error (MSE), normalized over noise variance, of reconstruction gives performance measure in the presence of noise. Our proposed Contour PSF has low MSE compared to other PSFs indicating high performance.
4.6 Proposed Contour-based PSF

We propose a contour-based PSF with high performance and validated by above mentioned metrics. We make following observations of characteristic of a high performing PSF:

1. Contain all directional filters to capture textural frequencies at all angles.

2. High contrast (or binary) to compensate for limited bit depth of sensor pixels.

3. Low support (or spatially sparse) to minimize the DC component of PSF’s Fourier transform.

In graphics, perlin noise [89] is popular technique to produce random landscape textures and contours. Such landscape contours contain a set of randomly oriented curves that can be used as directional filters. To get high contrast and low support, we apply canny edge detection to a generated perlin noise. Such generation of PSF is a good candidate for lensless imaging and have the above mentioned characteristics. Illustration of generating a contour PSF is shown in Fig. 6.

High performance of our proposed contour-based PSF is validated, in simulation, by the above mentioned metrics as shown in Fig. 5(e,f,g). We also compare our
Figure 7: Comparison with caustic pattern produced by 0.5° diffuser. Highest contrast pattern with diffuser was achieved at 16 mm. Our proposed Contour PSF has lower MSE compared to diffuser caustic and hence will perform better.

experimentally implemented PSF with DiffuserCam [18] PSF as shown in Fig. 7. Our PSF has lower theoretical MSE [88]. We did not experimentally compare reconstructions with the diffuser due to the larger camera thickness required for the diffuser caustic to be in focus.

4.7 Prototype

We generated a contour-based PSF with 14% sparsity from perlin noise [89]. Using the phase mask optimization mentioned in the previous section, we generated phase mask to form the PSF at sensor to mask distance of 1.95 mm with wavelength 532 nm and substrate refractive index of 1.5. The phase mask was fabricated using two-photon lithography 3D printer (Photonic Professional GT, Nanoscribe GmbH [90]). The PSF design, optimized phase mask profile and comparison of PSF from 3D printed phase mask is shown in Fig. 8.

Calibration. There can be discrepancies between the physically implemented PSFs and the target engineered PSFs due to phase mask height discretization and
fabrication inaccuracies. Hence, we experimentally capture the PSFs and use these PSFs for our computation. We approximate a point source by back illuminating a pinhole aperture and capturing sensor data by placing prototypes at different desired depths.

![Figure 8](image)

Figure 8: (a) Phase mask design. (b) Comparison between target PSF and captured PSF using the designed and implemented phase mask.

4.8 Reconstruction Algorithm

Diffractive lensless camera follows a convolutional forward model as discussed in section 4.3.4. The scene image can be reconstructed by deconvolving the PSF from the sensor measurement. We add Tikhonov regularization to the deconvolution to be ro-
bust to measurement noise and avoid large amplification of noise in the reconstruction. In effect, we solve the following minimization problem:

\[
\hat{i} = \min_i \frac{1}{2} \| b - p * i \|^2_2 + \frac{\gamma}{2} \| i \|^2_F ,
\]

(4.6)

where, \( b \) is the captured sensor measurement, \( p \) is the PSF at the scene depth, \( \gamma \) is the weighting of the regularization, and \( i \) is the image to be reconstructed. ‘\( * \)’ denotes the convolution operator and \( \| \cdot \|^2_F \) denotes the Frobenius norm. The above minimization problem has a closed form solution given by Wiener deconvolution, that can be computed in real time due to the use of the Fast Fourier Transform (FFT) algorithm. The closed form solution of reconstruction is given by:

\[
\hat{i} = F^{-1} \left( \frac{(F(p))^* \circ F(b)}{|F(p)|^2 + \gamma} \right),
\]

(4.7)

where, \( F(\cdot) \) is the fourier transform operator, \( (\cdot)^* \) is the complex conjugate operator and ‘\( \circ \)’ represents hadamard product.

### 4.9 Experimental Results

#### 4.9.1 Resolution

From geometry (Fig. 9), theoretical resolution of a lensless camera can be derived as:

\[
\text{Resolvable feature} = \frac{\text{Mask-Scene distance}}{\text{Mask-Sensor distance}} \cdot \text{(Pixel pitch)}
\]

(4.8)

From experimental testing using fluorescent USAF target, we find that using contour PSF achieves close to theoretical resolution as shown in Fig. 10. The pixel pitch of the camera is 4.8\( \mu m \).
Figure 9: Geometry of lensless camera. Theoretical resolution can be calculated from the similar triangles formed between pixel pitch and smallest achievable feature resolution.

Figure 10: Experimental evaluation of our camera’s resolution using fluorescent USAF target. The inserts are shown for line pairs with contrast close to 20% or higher.
4.9.2 Photography

Capture and reconstruction of scene in photography setting is shown in Figure 11.

4.9.3 Fluorescence Microscopy

Capture and reconstruction of scene in fluorescence microscopy setting is shown in Figure 11.

4.9.4 3D Imaging

Our recent and new (at the time of thesis submission) algorithm for 3D image reconstruction and the results can be found at [http://vivekboominathan.com/lensless-imaging](http://vivekboominathan.com/lensless-imaging). The developed algorithm is fast and stable in reconstruction high quality 3D images.
4.10 Discussion

In this chapter, we specifically focused on imaging as the application. However, we can use similar approach oriented to any given application. The next chapter, takes this further and shows application-specific miniature machine vision cameras.
Chapter 5

Miniature machine vision systems

Small, lightweight machine vision systems are critical for emerging technologies like implantables, robotics and automation, and the internet of things (IoT). While the electronic components of vision systems have witnessed dramatic miniaturization over the last decade, the corresponding optical elements have seen little reduction in size and weight. For electronics, miniaturization and performance have been achieved by using smaller transistors, distributed computing, and by replacing general-purpose central processing units with application-specific integrated circuits (ASIC). Here we propose an optical analog to ASICs for machine vision. Namely, we show that replacing general-purpose optical lenses with application-specific designed diffractive optical elements we can create designer point spread functions that enable miniature machine vision systems. As a proof-of-principle, we show two examples of diffractive-element-based miniature computer vision systems: (a) edge detection, and (b) head pose classification. In all cases, this approach produces an order of magnitude reduction in thickness, volume, cost, and weight compared to conventional lensed systems, suggesting that this “application-specific optical computing” provides a general framework for designing and building miniature machine vision systems.
5.1 Introduction

Cameras are becoming ubiquitous due to their expanding application space. Myriad emerging applications such as wearables, implantables, autonomous cars, robotics, internet of things (IoT), virtual/augmented reality, and human-computer interaction all make use of integrated cameras [8, 9, 10]. Traditionally, cameras were optimized primarily for imaging resolution; however, the needs of these new applications have created fundamentally different sets of performance requirements. Firstly, while image resolution is still an important attribute, visual information acquired from these sensors is no longer meant predominantly for human consumption, but rather for driving application-specific inferences. As a result, information content contained within the images is often more important than the aesthetic appeal. Secondly, as most emerging applications are mobile and untethered, camera systems must adhere to strict constraints in their size, weight, power, and cost (SWaPC) consumption [91, 92]. Current generation lens-based cameras are ill-equipped to simultaneously achieve both high resolution and low SWaPC due to design tradeoffs enforced for traditional lens-based cameras [16, 18, 17]. Specifically, Increasing resolution requires larger apertures [12, 93], larger lenses, and a corresponding increase in size and weight. Similarly, the increased resolution also results in increased data throughput and bandwidth needs that impact power and energy consumption.

A radical redesign of visual sensors is necessary to escape this trade-off imposed by the lens-based imaging design space. To redesign these systems, we draw inspiration from the field of computing that faced a similar challenge a few decades ago. While general purpose computing devices such as CPUs and DSPs were the cornerstone of computing (much like lens-based cameras are to imaging today), there were challenging SWaPC constraints when these general-purpose computing devices were
used in specific data-intensive applications. This led to the development of ASICs, GPUs, VPUs [94], and TPUs [95] that were application-specific computing platforms—designed, refined, and developed with application-specific constraints in mind. We believe that similar application-specific optical computing systems (ASOCS), that are designed and optimized for specific applications can break free from the resolution-SWaPC trade-off of lens-based designs and create new opportunities for visual sensing systems.

Traditional machine vision systems have a general purpose lens-based imaging system that captures images onto an array of photodetectors (a.k.a. an image sensor). Algorithms then extract features from this image and make inferences. While image sensors themselves are thin, light, compact and inexpensive, complete imaging systems still tend to be an order of magnitude worse than image sensors in terms of size, weight, thickness, and cost, primarily due to the need for physical lenses to focus an image and the associated optoelectronics. While lenses in camera modules have been miniaturized through advances in plastic lens [96] technology, this approach has reached its limit [10], and it is believed that further reductions in size, weight, and cost requires us to radically alter the design space for the camera. Recent works have shown that lens-free imaging systems [4, 18, 19, 17, 21, 22, 23, 7] provide a promising approach for achieving compact and light-weight cameras.

In this paper, we propose application-specific optical computing systems (ASOCS) realized by replacing the lens in a traditional camera with an application-specific designed diffractive optical elements placed close to the image sensor. Note that almost all computer vision and machine learning applications first capture an image and then compute application-dependent image features that are then fed as inputs to machine learning algorithms to derive inferences. Diffractive optical elements allow
us to compute these image features optically, on the image sensor chip while simultaneou
sly reducing the SWaPC footprint of these visual sensors by one or two orders of magni
tude. The particular image features that are useful depend upon the target applica
tion and are assumed known apriori at design time either through domain knowl
dge or through outputs of data-centric techniques. Given target image features
that need to be sensed/computed, we develop and demonstrate an algorithm for optimiz
ing the ASOCS and deriving optimal diffractive masks that can efficiently compute the target image features. The framework we develop is general in achieving an application-specific miniature vision system (Fig. 1), and we highlight the general-
ity of our framework by showcasing three radically different applications: (a) imaging, (b) edge feature extraction, and (c) head pose classification.

5.2 ASOCS design paradigm

ASOCS design consists of a target application-specific designed diffractive optical elements placed at a short distance (range of a few hundred microns to a couple of millimeters) away from a conventional CMOS or CCD sensor. Since both con-
tventional image sensors and diffractive optical elements are compact, light-weight, and inexpensive, the entire imaging system inherits these qualities. Besides, the de-
sign space of diffractive optical elements is a rich enough design space to achieve a wide variety of imaging and inference functionalities that are critical for varying target applications. Our design (Fig. 2) for a particular target application begins with identifying the features that are important for that target application (Step 1: Application-specific feature identification). Once target-specific features have been identified, these features are modeled as desired point-spread-functions for the imaging system and we solve an optimization problem to identify the optimal diffractive
Figure 1: Miniature Computer Vision System Design. (a) Traditional machine vision system has a general purpose optical imaging system followed by application-specific computation to achieve an application-specific goal. (b) In our miniature machine vision system, we design a front end application-specific optical computing systems (ASOCS) with thickness constraint followed by computational algorithms to achieve an application-specific goal. In this paper, we explore three different types of applications: imaging, oriented edge-extraction, and head pose classification.
mask pattern that can achieve direct feature measurement on the image sensor (Step 2: Phase mask optimization). This mask is then fabricated and placed at the appropriate distance from the image sensor – measurements from the image sensor now provide direct feature extraction which can then be processed by ML and AI algorithms for application-specific inferences as necessary (Step 3: Hardware realization and application Instantiation).

Among the three steps in our design (Fig. 2), steps 2 and 3 are application agnostic, while Step 1 provides all the application-specific criterion into the design.

Figure 2: ASOCS Design and Example applications. (a) Considering the application-specific goals and constraints, we design and develop the processes outlined in dashed lines. (b) Table laying out the goals, constraints, and ASOCS design components for three different applications: edge-extraction, and head pose classification.
5.2.1 Step 1: Application-dependent image feature selection

The first step is to identify image features that are critical for a particular application. This identification can be made in one of three ways.

1. Domain knowledge (See target application 1 for example): End users and experts in the target application domain usually have significant expertise regarding the selection of features that are best suited for specific applications, and we could rely on this expert knowledge to select image features for ASOCS.

2. General-purpose image features (See target application 2 for example): Over the last decade it has been shown that some general-purpose image features such as oriented gradients, Gabor features and first layer output of image-based neural networks such as AlexNet [97] are versatile and provide fairly robust and state of the art performance for a wide variety of target applications. We could use one of these as features for ASOCS.

3. Machine-learned optimal features adapted to target application (See target application 3 for example): When sufficient training data is available, data-dependent optimization and machine learning can be used to fine-tune and optimize image features that are optimal for specific target applications.

One of the three techniques laid out above will be used to identify the image features that are best suited for a target application. With an eye towards demonstrating all these three techniques, in the results section, we have shown one application where each of these three methods was used to identify the application-dependent target image features and design the respective point-spread-functions.
5.2.2 Step 2: Phase mask optimization

Once the image features that ASOCS must sense have been identified, the next step is to design the diffractive optical elements that are capable of directly capturing these image features using the image sensor chip.

Among the various diffraction masks, phase masks have proven to be versatile in realizing a variety of point-spread-functions [98, 99, 100, 101, 74, 102], with and without the assistance of lenses. Additionally, phase masks are highly light-efficient and hence operationally better suited for a range of illumination scenarios. Recognizing these beneficial properties of the phase mask, we build our ASOCS by optimizing and fabricating phase masks that realize our application-determined point-spread-functions.

The algorithmic details of phase mask optimization can be found in the earlier chapter under Section 4.4.

5.2.3 Step 3: Fabrication

Advancements in fabrication techniques have made it possible for physically realizing diffractive masks with quick turnaround times. In this paper, we use a recently developed 2-photon lithography 3D printing system [90] that allows for rapid prototyping of different phase masks without significant overhead preparation. With an optimized final phase mask design, fabrication can be scaled through the manufacturing pipeline such as photolithography and reactive-ion-etching processes.
5.3 Applications

In order to demonstrate the wide applicability and flexibility of ASOCS, here we demonstrate three completely different applications (a) high resolution imaging, (b) oriented edge extraction and (c) head pose classification. Together these three applications span across low-level, mid-level and high-level vision examples and demonstrate the generality of the approach. In our context, a low-level application recovers the scene image, a mid-level application extracts general purpose features that are useful for image analysis, and a high-level application performs inference tasks.

5.3.1 High resolution imaging

Recall that the miniature lensless imaging system discussed in Chapter 4 is low-level application of producing high-resolution images. For a lensless imaging system, the invertibility of PSF is crucial for reconstructing high-resolution images. The sensor measurement of a lensless system is the convolution of the mask’s PSF with the scene under view, and the scene image is reconstructed using a deconvolution process. Deconvolution involves inversion of the frequency spectrum, hence having large and almost flat magnitude spectrum is desirable in a PSF. Given that the PSF intensity is positive, we made the following observations in Section 4.6 that a high performing PSF: (a) contain all directional filters to capture textural frequencies at all angles, (b) high contrast (or binary) to compensate for limited bit depth of sensor pixels, and (c) low support (or spatially sparse) to minimize the DC component of PSF’s Fourier transform. Keeping these observations in mind, we proposed in Section 4.6 a heuristic-based contour PSF and showed that our proposed PSF is better (Fig. 5, Fig. 7) than previously proposed lensless imaging systems [4, 19, 7, 18]. Characteristics of our system is shown in Figure 3 and we achieve 5-10× reduction in size and weight.
**Traditional vs Miniature imaging system**

(a) Comparison of traditional vs our miniature imaging system. The ASOCS was designed with thickness constraint of 1.95 mm. (b) Proposed contour PSF. (c) Optimized phase mask profile for the proposed contour PSF.

**5.3.2 Oriented edge extraction**

Here we present a mid-level vision application of directly extracting useful features from the scene using ASOCS. Many image analysis techniques operate on textural
Figure 4: Miniature oriented-edge extraction system (a) Comparison of traditional system of extracting edge features vs our miniature system. The ASOCS was designed with thickness constraint of 4 mm. (b) Positive and negative components of Gabor filters of orientations 90deg and 45deg. (c) Optimized phase masks to implement the components of Gabor filters.

features extracted from the scene rather than on the images of the scene. One such class of useful features are oriented-edge features derived using oriented-edge filters. These edge filters are versatile and general-purpose, evident from the fact that they
Figure 5: *Oriented-edge extraction experiments.* (a) Experimentally realized filters and edge extractions at orientations 45deg and 90deg. The house image was captured by displaying on a monitor display while the face was captured directly. (b) Various filters that can be realized from the implemented PSFs using per-pixel subtraction.

appear in the first layer learned by most image-based neural networks [103, 104], and are also found in the V1 layer of the human visual cortex system [105]. Olhausen and Field [105] characterized these filters as Gabor filters and showed that the responses are maximally efficient in terms of information transfer.

To demonstrate the ASOCS ability to extract useful edge features, we optimized and implemented phase masks that realize Gabor filters as PSFs. Gabor filters, however, can not be achieved with just one mask, since they have both positive and negative values (Fig. 4b) while realizable PSF intensities are strictly non-negative. To overcome this limitation, we split the Gabor filters into positive and negative components and implemented them as non-negative PSFs using separate phase masks (Fig. 4c). The final edge features are then computed by per-pixel subtraction of outputs from the separate phase masks (Fig. 5a). We can also combine the implemented PSFs in different combinations to achieve various filters (Fig. 5b).
5.3.3 Head pose classification

Here we demonstrate the design of ASOCS for a high-level application of performing a classification task. The PSFs in the previous two applications of imaging and edge extraction were chosen using heuristic approaches. In this application, we optimize our machine vision system PSF using a data-driven end-to-end approach $^{101}$ $^{100}$, where we learn the ASOCS response together with the classifier from labeled data pairs. The learned PSFs are then implemented as phase masks using the near-field phase retrieval algorithm and then validated using a physical prototype.

![Diagram of Head pose classification]

Figure 6: *Head pose classification*. The ASOCS design was learned for thickness constraint of 1.95 mm. (b) Learned PSFs from CNN split into positive and negative components. (c) Optimized phase masks to realize the split PSFs. (d) Classification performance from implemented prototype.

Capitalizing on the recent success of convolutional neural networks (CNNs) in producing state of the art performance in many computer vision applications $^{97}$. 
we choose a CNN framework for machine learning the application-specific ASOCS response. A CNN architecture consists of a sequence of layers, where each layer performs convolution with learned filters followed by a non-linear activation function. ASOCS can replace the first layer of a CNN by performing the convolution optically. Previous works \cite{104,101,108} have optically replaced layers of CNN; however, ASOCS can achieve the same with strict thickness constraints.

![CNN architecture for Head Pose estimation](image)

**Figure 7**: CNN architecture and performance. (a) CNN architecture used for training the optical filters and testing performance. (b) Learned PSFs from CNN split into positive and negative components. (c) Optimized phase masks to realize the split PSFs. (d) Classification performance from implemented prototype.
As a specific example, we choose the application of head pose classification, where we classify the poses of faces as front-facing, left-facing, or right-facing. Using images from Pointing ‘04 database [109, 110], we train a convolutional-neural-network (CNN) architecture, where the first layer will be implemented using the phase mask (Fig. 6). The physically implemented prototype achieves an average performance of 90% in head pose classification. Characteristics of our prototype and performance are shown in Figure 6 and Figure 7.

5.3.4 Discussion

Overall this general framework to design light, efficient, and miniature lensless machine vision systems relies on two key steps. The first step is to optimize a PSF that is tailored for the given application and the second step is to optimize phase mask to attain the PSF at the desired camera thickness. Here we have demonstrated how this framework can be applied to three different applications: imaging, edge extraction, and head pose classification; however, the versatility of our framework allows the design of miniature machine vision systems for a variety of applications where size is a major constraint (e.g. micro-robotics, implantable electronics, and Internet of Things). An additional benefit of performing a layer of optical computing in the prior to capturing data with an image sensor is a the reduction of latency and power consumption. As described previously optical computing has benefits of high bandwidth, and inherently parallel processing, all potentially consuming zero power and happening at the speed of light [111]. The exact benefits for ASOCS for reduction in power and latency will likely be application dependent and will be an important aspect of future work. This aspect is currently not explored in this thesis but is an important element of future work. Additional future work should focus on the choice
of heuristics for the target PSF. In many cases the target PSF is selected based on the designer’s intuition, but more formal or automated methods could improve the choice of an “optimal” PSFs. In summary, this concept of improving machine vision performance by customizing the optics to capture information rather than images stands to improve the SWAPC of these systems in much the same way that computing systems benefit from the ASIC.
Chapter 6

Conclusion

Advancement in technology always involved miniaturization of components. For example, ability to make smaller and smaller transistors has led to portable and powerful computers. Every technological innovation is progressively made compact allowing for more innovations to take place.

Cameras is one of the important human inventions. Cameras has revolutionized the way we record and share information. Cameras gives visual context and has become an integral part of automation and artificial intelligence. Many of the emerging such as wearable and implantable electronics, micro-robotics and automation, and the internet of things (IoT) are driving the need for miniaturization of cameras. This is precisely the focus of the dissertation.

In Chapter 3 we demonstrated what is perhaps the world’s tiniest and lightest microscope, called FlatScope. FlatScope is a lensless microscope that is scarcely larger than an image sensor (roughly 0.2 grams in weight and less than 1 mm thick) and yet able to produce micron-resolution, high-frame-rate, 3D fluorescence movies covering a total volume of several cubic millimeters.

In Chapter 4 we presented a new lensless system response that can achieve high-resolution image reconstructions. We presented an algorithm to realize the system response with light-efficient phase masks and validated the performance with experiments.

In Chapter 5 we presented a generalized framework to design miniature lensless
cameras that is tailored for specific applications. With the framework, we hope to lay the groundwork for creating miniature computational cameras to augment low-power AI and meet the constraints of the future small-scale devices.
Bibliography


[46] M. Dierolf, A. Menzel, P. Thibault, P. Schneider, C. M. Kewish, R. Wepf, O. Bunk, and F. Pfeiffer, “Ptychographic X-ray computed tomography at the


2009.


