3D RECONSTRUCTION FROM CODED PLENOPTIC SAMPLING

by

Mingyuan Zhou

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science

Winter 2019

3D RECONSTRUCTION FROM CODED PLENOPTIC SAMPLING

by

Mingyuan Zhou

Approved: _____

Kathleen F. McCoy, Ph.D.
Chair of the Department of Computer and Information Science

Approved: _____

Levi T. Thompson, Ph.D.
Dean of the College of Engineering

Approved: _____

Douglas J. Doren, Ph.D.
Interim Vice Provost for Graduate and Professional Education

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Jingyi Yu, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Chandra Kambhamettu, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Li Liao, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

S. Susan Young, Ph.D.
Member of dissertation committee

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Prof. Jingyi Yu. His hardworking and passionate attitude inspires me a lot. He has taught me how to define a research problem and find a solution to it. I would like to thank him for his guidance, support and encouragement, as well as his valuable advice on both my research and career.

Besides, I would like to thank all the dissertation committee members, Prof. Chandra Kambhamettu, Prof. Li Liao and Dr. S. Susan Young, for their insightful comments on my research. I really appreciate Prof. Chandra Kambhamettu and Prof. Li Liao for their encouragement, great support and invaluable advice on my dissertation. I am also grateful to Dr. S. Susan Young for her contribution to the multi-spectral imaging related to this dissertation.

Moreover, I would like to thank all my colleagues in the Graphic and Imaging Lab at the University of Delaware, Yang Yang, Wei Yang, Zhong Li, Xinqing Guo, Qiaosong Wang, Can Chen, Nianyi Li, Yu Ji, Jinwei Ye and Haiting Lin, for their great support and for the fun we have enjoyed working together. More specifically, I am thankful to Yu Ji, Jinwei Ye, Nianyi Li and Haiting Lin for their contributions to various projects related to this dissertation.

I also would like to thank some students and researchers at ShanghaiTech University, Minye Wu, Shi Jin, Ruiyang Liu, Zhang Chen, Anpei Chen, Yinliang Zhang, Zhiru Shi, Wenguang Ma and Xuan Cao, for their collaboration on this dissertation.

I gratefully acknowledge the funding source that made my Ph.D work possible. My work is partially supported by the Army Research Office under the grant W911NF14-1-0338. Part of my research (Chapter 5) was conducted when I was an intern at Plex-VR.

Last but not least, I would like to express my deepest gratitude to my family. My parents have been helping and supporting me in every possible way, and no words can truly

express how grateful I am to them and how much I love them. I also want to thank my friends in UD and SIT. This dissertation would not be possible without their love and endless support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The plenoptic function describes a scene in terms of light rays, it is a 7-dimensional function with spectral, directional, spatial, and temporal variation. Traditional plenoptic sampling is acquired either by employing a standard plenoptic camera or a camera array, and the spatial-angular sampling can be potentially used to model 3D surface.

In this dissertation, I present three coded plenoptic sampling schemes, i.e., the rotational cross-slit (R-XSlit) plenoptic sampling, the wavelength coded plenoptic sampling, and the polarimetric plenoptic sampling. The additional coded sampling information, such as non-centric sampling, spectral sampling, and polarization sampling, are conducive to 3D reconstruction. Therefore, I also develop the corresponding 3D reconstruction framework for each of them.

First, I introduce the R-XSlit plenoptic sampling scheme by exploiting a special non-centric camera called the crossed-slit or XSlit camera. An XSlit camera acquires rays that simultaneously pass through two oblique slits. I show that instead of translating the camera as in the pinhole case, we can effectively sample the 4D plenoptic sampling by rotating individual or both slits while keeping the camera fixed, which makes the plenoptic sampling coded in the spatial-angular domain. The theoretical analysis shows that it provides denser spatial-angular sampling, which is beneficial for scene reconstruction and rendering. I develop a volumetric reconstruction scheme for scene reconstruction.

Second, I present two wavelength coded plenoptic sampling schemes in the visible and infrared spectrum respectively. I firstly design a compact system with lights and cameras arranged on concentric circles to acquire a concentric wavelength coded plenoptic sampling in the visible spectrum, the cameras on each ring capture images in a unique spectrum. I employ the Phong dichromatic model onto its plenoptic function for 3D reconstruction and spectral reflectance map estimation. Experiments show that our technique can achieve high

accuracy and robustness in geometry recovery. Moreover, I present an infrared wavelength coded plenoptic sampling and develop a hybrid sensing framework to efficiently achieve pose estimation and face reconstruction by exploiting the captured reflected infrared rays from human eyes.

Finally, I present a polarimetric plenoptic sampling framework for recovering 3D surfaces, the polarization of light is included in its plenoptic function. I employ a new analysis analogous to the optical flow to correlate the polarization radiance function with both surface normal and depth. The proposed framework effectively resolves the azimuth-zenith ambiguity by forming an over-determined system. Extensive experiments on both synthetic and real data demonstrate that the technique is capable of recovering extremely challenging glossy and textureless objects.

Chapter 1

INTRODUCTION

The plenoptic function, first introduced by Edward Adelson and James Bergen [3], describes radiance received along any direction arriving at any point in space, at any time and over any range of wavelength, which is a 7D function, as shown in Fig.1.1.(a). It encodes the 3D spatial and 2D directional information of a light ray. If we assume that the radiance remains consistent from point to point along the ray, one dimension of the plenoptic function will be redundant, thus making it possible to describe the light ray with a 4D function. A notable example for this is two-plane parameterization or 2PP, where a pair of parallel planes Π_{st} and Π_{uv} are given as priors in 3D space, and each ray is represented by its intersection with the planes as (s, t, u, v) [33], as shown in Fig.1.1.(b).

One of the most important tasks in image-based modeling and later computational photography and imaging is to conduct efficient sampling of the plenoptic function. Notable examples include capturing the scene using the plenoptic camera or the camera array. The former combines a lenticular array and a single high-resolution sensor with each lenslet emulating a pinhole camera, such as Lytro [1] and Raytrix [2]. Compared with the camera array, the plenoptic camera can sample more densely on the st dimension due to small microlenslet baselines, but at the sacrifice of the uv resolution, whereas the camera array facilitates a much wider baseline and can sample the angular dimension at a wider Field-of-View. At each camera location (s, t), it samples a uv slice corresponding to the image captured by the camera. Yu and McMillan [67] have shown that every sampled image corresponds to a 2D planar slice in the 4D plenoptic sampling. The camera array, in essence, samples the space using a sequence of 2D slices.

Previous application by utilizing the plenoptic sampling focused on refocused rendering and view interpolation. More recent approaches employ plenoptic sampling for 3D

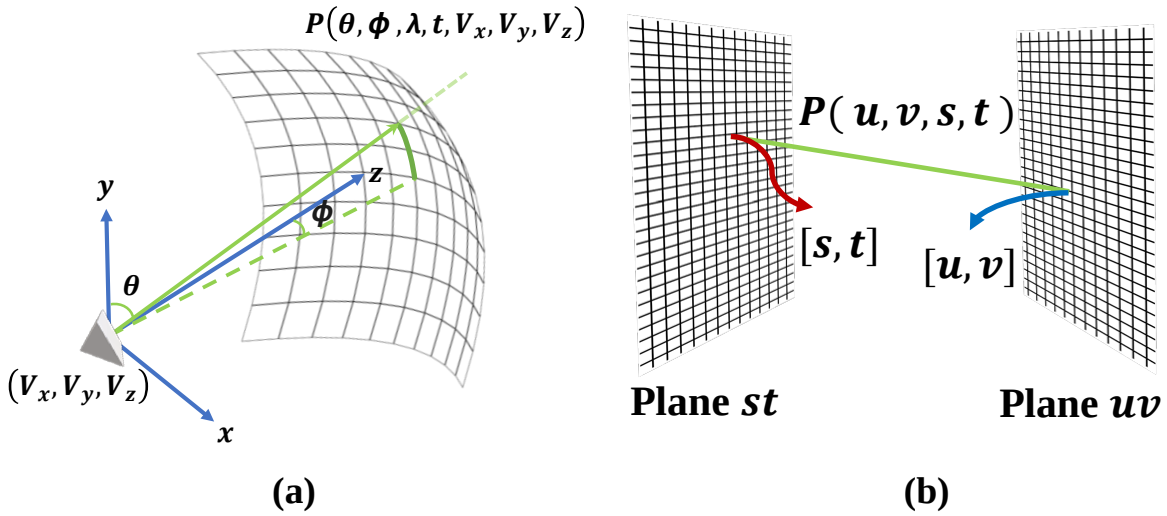


Figure 1.1: (a). The 7D plenoptic function. (b). The 4D light field plenoptic function.

reconstruction. [52, 58, 21]. For example, it is possible to directly use plenoptic cameras as 3D sensors [72]. However, 3D reconstruction from common plenoptic sampling is still a challenging problem, especially for specular, textureless and transparent surfaces.

1.1 Dissertation Statement

In this dissertation, I present three coded plenoptic sampling schemes, i.e., the rotational cross-slit (R-XSlit) plenoptic sampling, the wavelength coded plenoptic sampling and the polarimetric plenoptic sampling. Compared with the general plenoptic sampling scheme, the coded sampling information from each of them can be efficiently used for better 3D recovery and rendering.

First, I exploit the non-centric crossed-slit or XSlit camera model for acquiring the R-XSlit plenoptic sampling. In particular, the XSlit camera captures rays simultaneously passing through two oblique (neither parallel nor intersecting) slits in 3D space [75]. If the two slits are parallel to the 2PP, captured rays lie on a 2D planar surface in 4D ray space [64]. In fact, the pinhole camera can be viewed as a special XSlit camera where the two slits intersect. We adopt the design by Ye et al.[64] that relays two cylindrical lenses with slit apertures as the XSlit camera. To sample the plenoptic function, we rotate the XSlit camera along its optical axis, the R-XSlit plenoptic sampling is coded in the spatial-angular domain

by two rotational slits. We show the our R-XSlit sampling scheme provides substantial benefits. On the acquisition front, our new sampling scheme can achieve fixed-location acquisition. By rotating rather than translating the camera, we eliminate the need of building the camera array or moving the camera along the grid. On the reconstruction front, I show that the new sampling pattern enables more effective view synthesis and dynamic refocusing. Recall that the previous camera array samples uv slices at discrete st locations. Therefore, a new uv slice at an undersampled st location does not contain any samples and brute-force interpolation leads to severe ghosting or aliasing [70]. In contrast, I show under the R-XSlit sampling every perspective view will contain some minimum number of samples. The dense angular sampling is extremely helpful for 3D reconstruction and rendering.

Second, I present a novel concentric wavelength coded (CWC) plenoptic sampling method in the visible spectrum. We employ the Phong dichromatic reflectance model and integrate it into the plenoptic function to characterize the interface (specular) reflectance and the body (diffuse) reflectance. We design a concentric ring of cameras setup with a multi-spectral ring light to acquire CWC sampling. This setup imposes useful constraints on specularity variations that is used to robustly separate diffuse components from specular ones. Meanwhile, different lighting conditions can be captured under our multi-spectral ring light, then we can recover 3D surface shape and reflectance map by applying multi-spectral photometric reconstruction. Moreover, I propose another infrared wavelength coded (IRWC) plenoptic sampling scheme, and I develop a hybrid sensing camera array consisting of a pair of near infrared (NIR) cameras and a long-wave infrared (LWIR) camera, to sample infrared rays so as to facilitate pose estimation and 3D face reconstruction under poor lighting conditions.

Finally, I introduce a polarimetric plenoptic function with the polarization of the light ray. I synthesize a polarimetric camera array, where each sampling view has a specific polarization angle. Then, I derive a comprehensive theory that correlates the polarization radiance function with both surface normal and depth in terms of the transmitted radiance sinusoid and the polarization functions, and I develop a 3D reconstruction framework based on it and our polarimetric plenoptic sampling.

1.2 Contributions

This dissertation makes the following contributions.

Coded Plenoptic Sampling Scheme:

• I introduce four coded plenoptic sampling schemes.(R-XSlit plenoptic sampling, CWC plenoptic sampling, IRWC plenoptic sampling, and polarimetric plenoptic sampling).

Acquisition System Designs:

• I construct a R-XSlit camera system using a single camera with rotational Slits controlled by a programable motor. We align the two cylindrical lenses orthogonally using two lens tube. Each tube contains a rotation ring, with which I can control the rotation degree of each slit precisely.

• I construct a concentric multi-spectral camera array. Specifically, I mount a monochrome camera on a translation stage to uniformly translate the camera position on a 2D plane (i.e., the st plane). Then, I mount a tunable liquid crystal spectral filter in front of the camera to capture the scene under specified wavelengths. For the multi-spectral illumination part, I mount twelve LED chips onto a circle dodecagon frame, and then place twelve narrow-band spectral filters ranging from 450 nm to 670 nm with 20-nm step in front of the LED chips.

• I build an hybrid infrared sensing system consisting of a pair of NIR cameras and a LWIR camera. I strategically surround each NIR sensor with a ring of LED IR flashes to capture "bright-eyes" effect of the target, which can be used to accurately determine the face pose and geometry. The LWIR camera is used to capture potential targets as reliable sources. I design a control system to synchronize and control all sensors and the infrared lights.

• I establish a polarimetric sampling acquisition system through a translation rig and a polarization camera. The polarization camera can sample light rays with its polarizer array which is comprised of four different angled polarizers.

Algorithm Developments:

• I develop a volumetric reconstruction scheme applied on our R-XSlit plenoptic sampling for scene reconstruction. I discretize the scene into voxels, and then use the XSlit back-projection to map the voxels onto each XSlit view. The 3D embedded voxel graph is optimized by the graph-cut algorithm.

• I introduce a multi-spectral surface camera (MSS-Cam) by extending the classical surface camera (S-Cam) [68] with the Phong dichromatic model, i.e., each ray sampled in the MSS-Cam originated from the same 3D point, but at a different angle and with a specific spectrum.

• I propose a new appearance consistency metric applied on the MSS-Cam and a robust confidence metric for separating Lambertian and non-Lambertian points. I develop a specular removal scheme for non-Lambertian points and a multi-spectral photometric stereo technique for 3D reconstruction.

• I develop a 3D face reconstruction framework based on the IRWC plenoptic sampling. A modified cascaded CNN is represented for thermal facial landmarks detection in the framework.

• I derive a comprehensive theory that correlates the polarization radiance function with both surface normal and depth. Based on this derivation, I extend the shape-from-motion theory by viewing the plenoptic sampling as a moving camera and then derive a new formulation under our polarimetric plenoptic sampling for shape reconstruction.

1.3 Blueprint of the Dissertation

This dissertation is organized as follows:

Chapter 2 discusses general plenoptic sampling and reviews some kinds of works in 3D reconstruction which are highly related to this dissertation.

Chapter 3 introduces the R-XSlit plenoptic sampling scheme. It analyzes the sampling pattern and the blur kernel, and explores the epipolar geometry problem. Then, the rendering technique and 3D reconstruction method are discussed.

Chapter 4 presents two wavelength coded plenoptic sampling schemes. It introduces the MSS-Cam for the CWC plenoptic sampling and makes an analysis on it for 3D reconstruction. Then, it introduces the IRWC plenoptic sampling scheme and a 3D face reconstruction framework.

Chapter 5 introduces the polarimetric plenoptic sampling, a comprehensive theory that correlates the polarization radiance function with both surface normal and depth, and

then a new formulation can be derived under the polarimetric plenoptic sampling for 3D reconstruction.

Chapter 6 concludes this dissertation, discusses future work and lists some unresolved questions.

Chapter 2

BACKGROUND AND PREVIOUS WORKS

In this chapter, we review the background of the plenoptic sampling and some previous works in 3D reconstruction which are closely related to our work.

2.1 Plenoptic Sampling

Edward Adelson and James Bergen [3] first introduced the concept to computer vision and graphics via the 7D plenoptic function, which later became the foundation for image-based modeling and rendering. The plenoptic function expresses the image of a scene from all possible 3D viewing positions and 2D directions, but its high dimensionality prevents it from practical uses. Levoy and Hanranhan [33] introduced a practical light field representation using two-plane-parametrization or 2PP, where each plane describes a 2D subset, and the overall plenoptic function is 4D.

By far most commonly used devices for sampling plenoptic function include a moving hand-held camera or robotically controlled camera [42, 54], an 1D array of cameras [74](as used in capturing the bullet time effect in the film The Matrix), a dense array of cameras [54], and most recently hand-held plenoptic cameras [42] based on the lenslet array or coded apertures [56]. The MIT construct a camera array employing a grid of 64 1.3 megapixel usb webcams and the Stanford array is a two-dimensional grid composed of 128 1.3 megapixel Firewire cameras. With the help of registration, it is also possible to sample the plenoptic function by waving a camera in 3D space. Nearly all existing solutions use (e.g., in a camera array) or emulate (as in a lenslet array) perspective cameras as the main acquisition apparatus. The perspective camera sampling theory has been well studied in both spatial and frequency domains [14].

While the pinhole camera has been the most common device for imaging including acquiring plenoptic sampling, a tendency towards adopting the non-centric camera has gradually emerged. Classic examples include the pushbroom camera [67] which collects rays along parallel planes from points along a linear trajectory and the crossed-slit camera which collects all rays passing through two oblique lines. The General Linear Camera framework [66] discovers that rays collected by both pushbroom and XSlit cameras, together with classical perspective and orthographic cameras, correspond to 2D planar slices in the 4D plenoptic space.

2.2 3D Reconstruction

2.2.1 XSlit Stereo

An XSlit camera captures rays that simultaneously pass through two oblique slits in 3D space. A translational XSlit stereo model was introduced by Feldman et al.[18], which presents that the valid stereo pairs with purely horizontal parallax can be formed by translating an XSlit camera along one of the two slits. Seitz [49] and Pajdla [44] independently classified all possible stereo pairs according to their epipolar geometry. Their results show that apart from perspective camera pairs whose epipolar geometry is a plane, there exists another two kinds of epipolar geometry: hyperboloids and hyperbolic-paraboloids, both corresponding to double ruled surfaces. Ye et al.[64] developed a rotational XSlit stereo matching based on hyperboloids and validated Seitz's theory. Instead of translating the XSlit cameras, Ye et al.form valid stereo pairs by fixing sensor locations but switching the slits' directions and they show a theoretical analysis to characterize their rotational XSlit epipolar geometry.

2.2.2 Multi-Spectral Photometric Stereo

In the 1980's, Woodham [61] introduced the Photometric Stereo (PS), an extension of shape-from-shading [26] to recover surface normal by placing the object under different lighting conditions. In the beginning, most PS methods focus on the reconstruction of static objects. Later, some approaches extend the general PS for the dynamic scene by employing temporal and spectral multiplexing.

PS approaches with temporal multiplexing employ rapidly alternating lights. Vlasic et al.[57] built a large dome with 1200 individually controllable light sources to provide a series of spherical lighting and used high frame rate camera to capture the normal map. Ma et al.[35] captured the normal map using time-multiplexed illumination consisting of structured and polarized lights. These methods need to cope with image misalignment since each frame is captured under different lighting conditions at different times.

By contrast, the spectral multiplexing methods [24, 62, 5, 23, 30] use different colored lights to capture different lighting conditions in a single snapshot, thus making PS possible to conduct per-frame photometric reconstruction. In practice, most of them impose a monochromaticity constraint on the objects since they cannot directly estimate the surface normal (2 unknowns) and albedo (3 unknowns) for each pixel just from three color channels. Anderson et al.[5] relaxed the monochromaticity constraint by an assumption of multiple piecewise constant chromaticities. Furthermore, Fyffe et al.[19] used a beam splitter and two Dolby dichroic filters to obtain six-channel photographs, relaxing the chromaticity restriction, however, the calibration results in significant bias in the reconstruction since variation in the spectral reflectance cannot be represented by a 3-dimensional basis.

2.2.3 Shape from Polarization

There has been emerging interest on analyzing the polarization state of reflected light to infer surface geometry. When unpolarized light is reflected from a dielectric surface, it becomes partially polarized. Previous works [38, 40, 28, 8, 37] use the phase angle information from the transmitted radiance sinusoid function to estimate the azimuth angle of surface normal, and employ the polarization angle derived from Fresnel reflectance theory to determine the zenith angle. However, the problems are inherently ill-posed due to azimuth-zenith ambiguities.

To resolve the azimuth-zenith ambiguities, single view approaches [60, 39] employ additional constraints such as surface geometry priors and lighting assumptions. And [7] demonstrated that diffuse polarization from the dielectric surface does not exhibit zenith ambiguity. Besides, multi-view approaches can also help mitigate ambiguity. Rahmann et

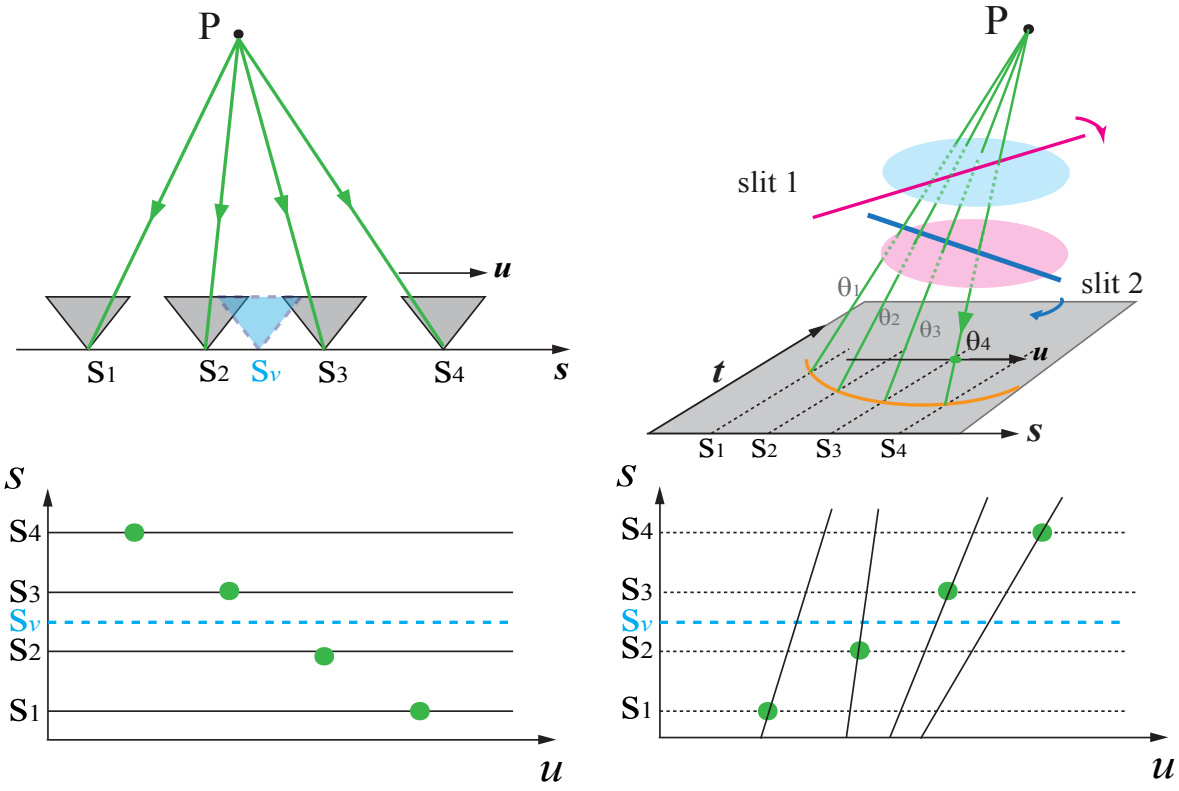
al.[47] and Atkinson et al.[6] constructed a stereo setup integrating coarse depth estimation with polarization cues. [9, 48] obtained correspondences from different viewpoints and resorted to simple geometric shape priors to address ambiguity. Cui et al.[16] proposed a polarimetric multi-view stereo that combines shape from polarization and epipolar constraints. Specifically, they utilized the classical structure-from-motion and multi-view stereo to obtain an initial shape estimation which helps to remove ambiguity.

Chapter 3

SCENE RECONSTRUCTION FROM ROTATIONAL CROSSED-SLIT PLENOPTIC SAMPLING

In this chapter, we present a novel plenoptic sampling scheme via rotating XSlit and its applications in scene reconstruction. Specifically, we exploit the non-centric crossed-slit or XSlit camera to sample rays. An XSlit camera captures rays that simultaneously pass through two oblique (neither parallel nor intersecting) slits in 3D space [75]. If the two slits are parallel to the 2PP, the captured rays lie on a 2D planar surface in the 4D ray space [64]. In fact, the pinhole camera can be viewed as a special XSlit camera where the two slits intersect. Although XSlit geometry has been thoroughly studied [75, 67], recently practical designs [64] have began to use it for computer vision tasks such as scene understanding and reconstruction [64, 65].

We adopt the design by Ye et al.[64] that relays two cylindrical lenses with slit apertures as the XSlit camera. To sample rays, our approach is to rotate the XSlit camera along its optical axis, and we show that the resulting Rotational XSlit (or RXSlit) sampling scheme provides substantial benefits. On the acquisition front, our new sampling scheme can achieve "fixed-location" plenoptic sampling by rotating rather than translating the camera. On the reconstruction front, we show that the new sampling pattern enables more effective view synthesis and dynamic refocusing. Recall that the previous camera array samples uv slices at discrete st locations. Therefore, a new uv slice at an undersampled $s't'$ location does not contain any samples, and brute-force interpolation leads to severe ghosting or aliasing [70], as shown in Fig. 3.1(a). In contrast, we show that the rotational XSlit sampling scheme ensures that every perspective view contains a minimum number of samples, as presented in Fig. 3.1(b).



(a) Perspective camera array (b)Rotational XSlit camera

Figure 3.1: The sampling pattern using a pinhole camera array (a) and using a rotational XSlit camera (b). A new perspective view (blue line) may not contain any sample in the pinhole case but is guaranteed to contain samples in the XSlit case.

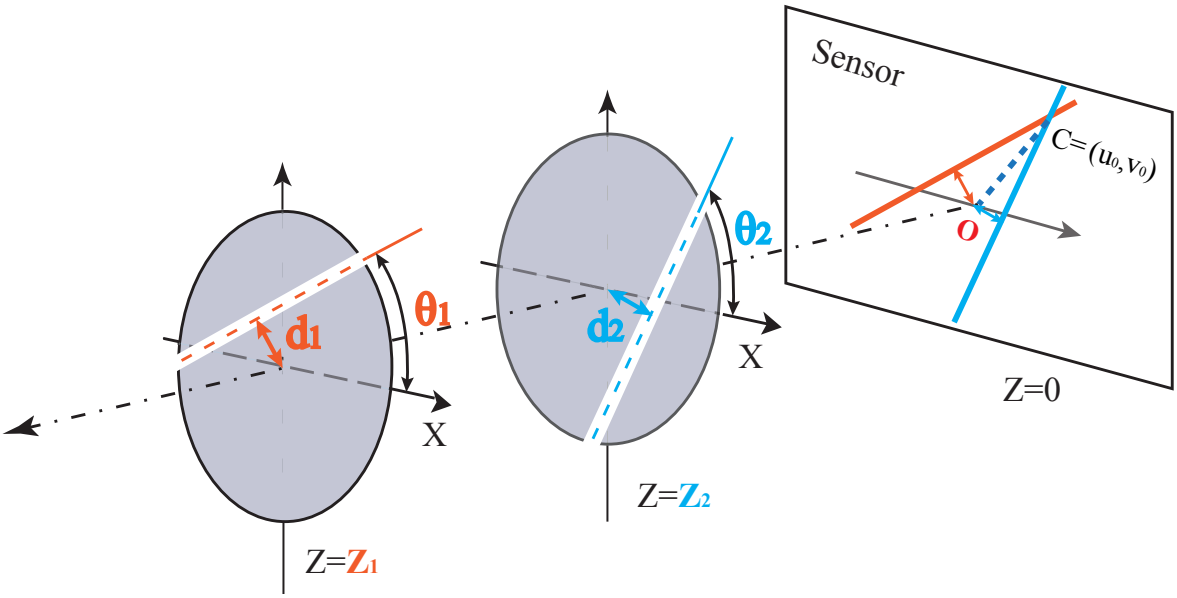


Figure 3.2: Illustration of our XSlit camera models. The center ray drifts off the image center.

We further validate our approach on using the R-XSlit sampling for dynamic scene refocusing and volumetric reconstruction. For 3D reconstruction, we discretize the scene into voxels and apply XSlit back-projection to map the voxels onto each XSlit view and optimize the the 3D embedded voxel graph by the graph-cut algorithm. For scene refocusing, analogous to refocusing with a camera array, we specify a proxy geometry plane and then project all XSlit views onto the plane. The refocused results exhibit some unique effects: defocused blurs become more severe on pixels farther away from the image center. This leads to a novel refocusing effect that we call "Conic Blur". Experiments on both synthetic and real scenes show that our methods are robust and reliable.

3.1 Plenoptic Sampling via Rotating XSlit

In this section, we discuss how to acquire the plenoptic sampling via rotating XSlit. An XSlit camera collects rays that simultaneously pass through two oblique (neither parallel nor coplanar) slits in 3D space [43, 75, 67]. We first adopt the two-plane parametrization [33] for its simplicity. Specifically, we choose two planes Π_{uv} and Π_{st} parallel to both slits but containing neither slits.

We will also use position-direction parametrization $[u, v, \sigma, \tau]$ where $\sigma = s - u$ and $\tau = t - v$ to simplify the analysis. We choose Π_{uv} as the default image (sensor) plane so that (u, v) can directly represent the pixel coordinate and $(\sigma, \tau, 1)$ can be viewed as the direction of the ray. Ye et al.[65] assumed that the origin of the coordinate system is the intersection point of the two slits' projected lines on Π_{uv}. We explore a more general case, i.e., the origin biases that intersection point and two slits rotate along z-axis. We assume that the two slits, l_1 and l_2, lie at $z = Z_1$ and $z = Z_2$ with angles θ_1 and θ_2 w.r.t.the x-axis, and the distance between their projected lines on Π_{uv} and the origin point are d_1 and d_2, where $Z_1 > Z_2 > 0$ and $\theta_1 \neq \theta_2$, as shown in Fig. 3.2. Each XSlit camera can be represented as $\mathcal{C}(Z_1, Z_2, \theta_1, \theta_2, d_1, d_2)$. We applied this notation for the sampling by changing θ_1 and/or θ_2. Thus, each pixel (u, v) in $\mathcal{C}$ maps to a ray with direction $(\sigma, \tau, 1)$, and there must exist some a_1 and a_2 so that:

$$\begin{cases} u + Z_1\sigma + \dfrac{d_1}{\sin\theta_1} = a_1\cos\theta_1; \quad v + Z_1\tau = a_1\sin\theta_1 \\ u + Z_2\sigma + \dfrac{d_2}{\sin\theta_2} = a_2\cos\theta_2; \quad v + Z_2\tau = a_2\sin\theta_2 \end{cases} \tag{3.1}$$

Eliminating a_1 and a_2, we obtain:

$$\begin{cases} (u + Z_1\sigma)\sin\theta_1 + d_1 = (v + Z_1\tau)\cos\theta_1 \\ (u + Z_2\sigma)\sin\theta_2 + d_2 = (v + Z_2\tau)\cos\theta_2 \end{cases} \tag{3.2}$$

So that we obtain two linear constraints as:

$$\begin{cases} \dfrac{Z_1\cos\theta_1}{Z_2\cos\theta_2} = \dfrac{(u + Z_1\sigma)\sin\theta_1 + d_1 - v\cos\theta_1}{(u + Z_2\sigma)\sin\theta_2 + d_2 - v\cos\theta_2} \\ \dfrac{Z_1\sin\theta_1}{Z_2\sin\theta_2} = \dfrac{(v + Z_1\tau)\cos\theta_1 - u\sin\theta_1 - d_1}{(v + Z_2\tau)\cos\theta_2 - u\sin\theta_2 - d_2} \end{cases} \tag{3.3}$$

Then, the $[\sigma, \tau]$ can be derived as:

$$\begin{cases} \sigma = (Au + Bv + F)/E \\ \tau = (Cu + Dv + G)/E \end{cases} \tag{3.4}$$

where

$$
\begin{aligned}
A &= Z_2 \cos\theta_2 \sin\theta_1 - Z_1 \cos\theta_1 \sin\theta_2, \\
B &= (Z_1 - Z_2) \cos\theta_1 \cos\theta_2, \\
C &= (Z_2 - Z_1) \sin\theta_1 \sin\theta_2, \\
D &= Z_1 \cos\theta_2 \sin\theta_1 - Z_2 \cos\theta_1 \sin\theta_2, \\
E &= Z_1 Z_2 \sin(\theta_2 - \theta_1), \\
F &= (d_1 \cdot Z_2) \cos\theta_2 - (d_2 \cdot Z_1) \cos\theta_1, \\
G &= (d_1 \cdot Z_2) \sin\theta_2 - (d_2 \cdot Z_1) \sin\theta_1.
\end{aligned}
$$

To sample rays via rotating XSlit , we simultaneously rotate both slits while maintaining their relative angle. To simplify our model, we assume that POX-Slit camera where the angle between the two slits kept at 90 degrees [65] captured two such images through rotating the camera by 90 degrees to conduct stereo matching. We characterize ray sampling pattern when exhausting all possible rotation angles and denote the plenoptic sampling scheme as $\mathcal{C}(Z_1, Z_2, \theta + 90^o, \theta, d_1, d_2)$ (abbreviated as $\mathcal{C}_\theta$ for simplicity), for all θ. A major advantage of this sampling scheme is that we can rotate the XSlit camera or the XSlits lens set as a unit instead of rotating individual slit. Then the Eqn. 3.4 can be simplified as:

$$
\begin{cases}
\sigma = (A'u + B'v + F')/E' \\
\tau = (B'u + D'v + G')/E'
\end{cases}
\tag{3.5}
$$

where

$$
\begin{aligned}
A' &= Z_2 \cos^2\theta + Z_1 \sin^2\theta, \\
B' &= \frac{(Z_2 - Z_1)}{2} \sin(2\theta), \\
D' &= Z_1 \cos^2\theta + Z_2 \sin^2\theta, \\
E' &= -Z_1 Z_2, \\
F' &= (d_1 \cdot Z_2) \cos\theta + (d_2 \cdot Z_1) \sin\theta, \\
G' &= (d_1 \cdot Z_2) \sin\theta - (d_2 \cdot Z_1) \cos\theta.
\end{aligned}
$$

3.1.1 Sampling Pattern

To analyze the plenoptic sampling pattern, we fix pixel $p = (u_0, v_0)$ on the sensor plane Π_{uv} and then analyze the sampled rays that pass through p. Specifically, we characterize the sampling function with respect to Π_{st}, i.e., the plane recording the angular information of all rays when rotating the camera. We assume that l_1 and l_2 have an infinite length, and then compute (σ, τ) for (u_0, v_0) in camera $\mathcal{C}_\theta$ with Eqn. 3.5. Since $s = \sigma + u$, $t = \tau + v$, we prove that the collect rays form a ring on the st plane as:

$$\begin{cases} s = (A'u_0 + B'v_0 + F')/E' + u_0 \\ t = (B'u_0 + D'v_0 + G')/E' + v_0 \end{cases} \tag{3.6}$$

Then, Eqn. 3.6 becomes:

$$\begin{cases} s = (1 - \frac{1}{Z_1}\cos^2\theta - \frac{1}{Z_2}\sin^2\theta)u_0 + (\frac{1}{2Z_2} - \frac{1}{2Z_1})\sin(2\theta)v_0 - \frac{d_1}{Z_1}\cos\theta - \frac{d_2}{Z_2}\sin\theta \\ t = (1 - \frac{1}{Z_2}\cos^2\theta - \frac{1}{Z_1}\sin^2\theta)v_0 + (\frac{1}{2Z_2} - \frac{1}{2Z_1})\sin(2\theta)u_0 - \frac{d_1}{Z_1}\sin\theta + \frac{d_2}{Z_2}\cos\theta \end{cases} \tag{3.7}$$

So the $[s, t]$ can be derived as:

$$\begin{aligned} s &= c_s - r_{\alpha_s}\cos(\theta - \alpha_s) + r_{\beta_s}\cos(2\theta - \beta_s) \\ t &= c_t - r_{\alpha_s}\sin(\theta - \alpha_s) + r_{\beta_s}\sin(2\theta - \beta_s) \end{aligned} \tag{3.8}$$

where

$$\begin{aligned} c_s &= u_0(1 - \frac{1}{2Z_1} - \frac{1}{2Z_2}) \\ c_t &= v_0(1 - \frac{1}{2Z_1} - \frac{1}{2Z_2}), \\ r_{\alpha_s} &= \sqrt{(\frac{d_1}{Z_1})^2 + (\frac{d_2}{Z_2})^2}, \\ r_{\beta_s} &= \sqrt{u_0^2 + v_0^2}(\frac{1}{2Z_2} - \frac{1}{2Z_1}), \\ \alpha_s &= \arctan\frac{d_2 Z_1}{d_1 Z_2}, \\ \beta_s &= \arctan(v_0/u_0) \end{aligned}$$

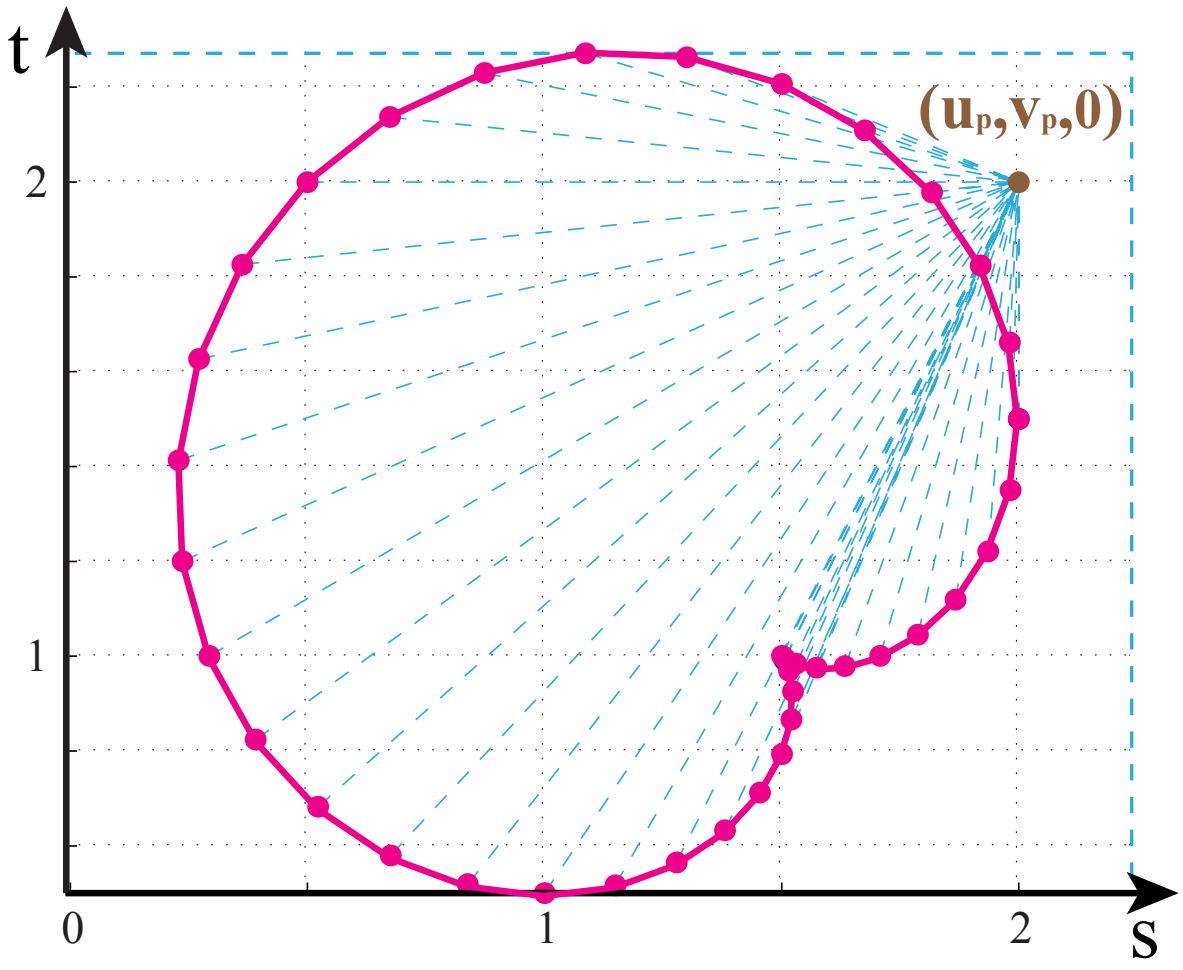


Figure 3.3: The (s,t) locus of (u_p, v_p) when varying θ from 0 to 2π.

This reveals that all (s,t) lie on a Limacon of Pascal curve, as shown in Fig. 3.3. It is important to note that when $d_1 = d_2 = 0$ the Limacon of Pascal will degrade to a circle.

Compared with 4D plenoptic sampling scheme (such as light field (LF)) using a projective camera array, such rotation-based sampling scheme has a few advantages. Firstly, our scheme can acquire many more angular samples which correspond to the number of different rotation angles whereas the angular resolution in the projective camera array corresponds to the number of cameras. What is more important is that it provides much denser angular sampling. In the camera array case, its density depends on the space between cameras, making it difficult to keep the baseline small enough to avoid undersampling or aliasing. In contrast, in the rotational XSlit, we can make the rotation step very small to acquire a highly dense rays sampling. Although the LF camera can potentially achieve the same results using tailored optical units, e.g., a microlenslet array, our sampling scheme does not require any special optical device. Secondly, it is much easier to rotate the slits than to build a camera array or translation stage to control the camera.

Fig. 3.1 shows the sampling differences between the traditional perspective camera array and our rotational XSlit camera. For the former, we show a 2D slice su from a 4D plenoptic sampling captured by conventional camera/lenticular array. Under this sampling, each image captured by a camera maps to a 2D parallel slice. Since the space between adjacent slices are "empty", any new perspective view, corresponding to a slice in between, will not contain any sampled ray and traditional approaches rely on geometry-guided ray interpolation [29]. For the latter, the plenoptic sampling with our rotational XSlit camera setup samples the space in a different way: each XSlit camera maps to a 2D slice [67] but under the rotational setup and the recorded slices are not axis-aligned in the 4D ray space. As a result, if we render a new perspective view (2D slice), it is guaranteed to intersect with the sampled XSlit slices and contain a minimal number of ray samples.

3.1.2 Blur Kernel

Given our R-XSlit sampling captured by $\mathcal{C}_{\theta_i}$, $i = 1, ..., N$, and a 3D point $X = (x_0, y_0, z_0)$ in the world. For each ray $[u, v, \sigma, \tau]$ passing through X, there exist some a_z that

satisfies:

$$[u, v, 0] + a_z[\sigma, \tau, 1] = [x, y, z] \tag{3.9}$$

By eliminating a_z, we have:

$$\begin{cases} u = x_0 - z_0\sigma \\ v = y_0 - z_0\tau \end{cases} \tag{3.10}$$

Combining above Eqn. 3.10 with Eqn. 3.5, we can derive two linear constraints as follows:

$$\begin{cases} u = x_0 - z_0((-\frac{\cos^2\theta}{Z_1} - \frac{\sin^2\theta}{Z_2})u + (\frac{1}{2Z_2} - \frac{1}{2Z_1})sin(2\theta)v - \frac{d_1}{Z_1}\cos\theta - \frac{d_2}{Z_2}\sin\theta) \\ v = y_0 - z_0((-\frac{\cos^2\theta}{Z_2} - \frac{\sin^2\theta}{Z_1})v + (\frac{1}{2Z_2} - \frac{1}{2Z_1})sin(2\theta)u - \frac{d_1}{Z_1}\sin\theta + \frac{d_2}{Z_2}\cos\theta) \end{cases} \tag{3.11}$$

Then, we have:

$$\begin{cases} \frac{(\frac{1}{2Z_2} - \frac{1}{2Z_1})\sin(2\theta)z_0}{1 - (\frac{\cos^2\theta}{Z_2} + \frac{\sin^2\theta}{Z_1})z_0} = \frac{x_0 - u + (\frac{\cos^2\theta}{Z_1} + \frac{\sin^2\theta}{Z_2})uz_0 + (\frac{d_1}{Z_1}\cos\theta + \frac{d_2}{Z_2}\sin\theta)z_0}{y_0 - (\frac{1}{2Z_2} - \frac{1}{2Z_1})\sin(2\theta)uz_0 + (\frac{d_1}{Z_1}\sin\theta - \frac{d_2}{Z_2}\cos\theta)z_0} \\ \frac{1 + (\frac{\cos^2\theta}{Z_1} + \frac{\sin^2\theta}{Z_2})z_0}{(\frac{1}{2Z_2} - \frac{1}{2Z_1})\sin(2\theta)z_0} = \frac{x_0 - (\frac{1}{2Z_2} - \frac{1}{2Z_1})\sin(2\theta)vz_0 + (\frac{d_1}{Z_1}\cos\theta + \frac{d_2}{Z_2}\sin\theta)z_0}{y_0 - v + (\frac{\cos^2\theta}{Z_2} + \frac{\sin^2\theta}{Z_1})vz_0 + (\frac{d_1}{Z_1}\sin\theta - \frac{d_2}{Z_2}\cos\theta)z_0} \end{cases} \tag{3.12}$$

The $[u, v]$ can be solved w.r.t. X as:

$$\begin{aligned} u &= c_u + r_{\alpha_b}\cos(\theta + \alpha_b) + r_{\beta_b}\cos(2\theta - \beta_b) \\ v &= c_v + r_{\alpha_b}\sin(\theta + \alpha_b) + r_{\beta_b}\sin(2\theta - \beta_b) \end{aligned} \tag{3.13}$$

where

$$\begin{aligned} c_u &= -\frac{x_0}{2}(\frac{Z_1}{z_0 - Z_1} + \frac{Z_2}{z_0 - Z_2}) \\ c_v &= -\frac{y_0}{2}(\frac{Z_1}{z_0 - Z_1} + \frac{Z_2}{z_0 - Z_2}), \\ r_{\alpha_b} &= z_0\sqrt{(\frac{d_1}{z_0 - Z_1})^2 + (\frac{d_2}{z_0 - Z_2})^2}, \\ r_{\beta_b} &= \frac{\sqrt{x_0^2 + y_0^2}}{2}(\frac{Z_2}{z_0 - Z_2} - \frac{Z_1}{z_0 - Z_1}), \\ \alpha_b &= \arctan\frac{d_2(z_0 - Z_1)}{d_1(z_0 - Z_2)}, \\ \beta_b &= \arctan(y_0/x_0) \end{aligned}$$

Refocusing with R-XSlit camera Refocusing with camera array

Figure 3.4: Refocusing rendering comparison between the R-XSlit sampling and regular sampling.

We set out to analyze the shape and size of blur kernel by finding the pattern of all the projections of X on a plane Π_f at $z = f$ parallel to the sensor plane. We compute the projection (u_f, v_f) as:

$$\begin{cases} u_f &= (1 - f/z_0)u + x_0 f/z_0 \\ v_f &= (1 - f/z_0)v + y_0 f/z_0 \end{cases} \tag{3.14}$$

According to Eqn. 3.13 and 3.14, the projection trajectory of X on plane Π_f is a Limacon of Pascal. The kernel size depends on the spatial location of X. Getting Closer to the center optical axis or further away from the slits will result in a smaller blur kernel size. This dependency of blur size on depth and spatial center is consistent with our vision habit:: we focus on an important object and make it centered in the view. Previous studies in biology [17, 45] have shown that human eyes capture a much higher resolution near the center of the retina than near the boundary. Similarly, in our R-XSlit plenoptic sampling scheme, rays are much more densely sampled (angularly) near the center. Consequently, when we conduct the refocusing via ray blending, our uneven ray sampling leads to non-uniform refocusing. Such a phenomena is very common to the human perception system [41, 13] and [10, 4] have already explored this "Conic Blur" property in video extrapolation. Therefore, we believe that the refocusing rendering from the R-XSlit will naturally conform with our vision system.

3.1.3 Epipolar Geometry Existency

The image sequence captured by rotating both slits generally does not form valid epipolar geometry. Ye et al.[65] have shown that the necessary and sufficient condition for two XSlit cameras to form valid epipolar geometry is when the directions of the two slits get switched, *i.e.* between $\mathcal{C}(Z_1, Z_2, 0, 90^o, 0, 0)$ and $\mathcal{C}(Z_1, Z_2, 90^o, 0, 0, 0)$. In the special POX-Slit case, where the two slits are perpendicular, every image in the captured sequence can form epipolar geometry together with the other in the sequence, i.e., the one whose slit directions are flipped, if we rotate the camera to cover 360 degrees. Finally, it is worth noting that even for cases without valid epipolar geometry, we can still conduct efficient volumetric reconstruction.

3.2 Scene Reconstruction

In this section, we demonstrate applications of our rotational XSlit plenoptic sampling scheme.

3.2.1 Volumetric Reconstruction

Recall that the R-XSlit camera does not have epipolar geometry across all views. The only case where epipolar pairs exist is when $d_1 = d_2 = 0$. Such a sampling scheme can be viewed as multiple stereo pairs despite that no uniform epipolar geometry exists across all pairs. In this case, we can adopt the volumetric reconstruction scheme for both 3D recovery and rendering.

The problem of reconstruction can be formulated as a variation of the fundamental space carving framework by Kutulakos and Seitz [31], which leverages a set of N perspective input camera views to recover a 3D volumetric representation of the scene. In classical volumetric reconstruction methods, it first discretize the scene into voxels coherent with the resolution of the input image. In our case, we first positions a virtual perspective camera whose Center-of-Projection lies at $(0, 0, Z)$, where $Z^{=}(Z_1 + Z_2)/2$ with the size of its view

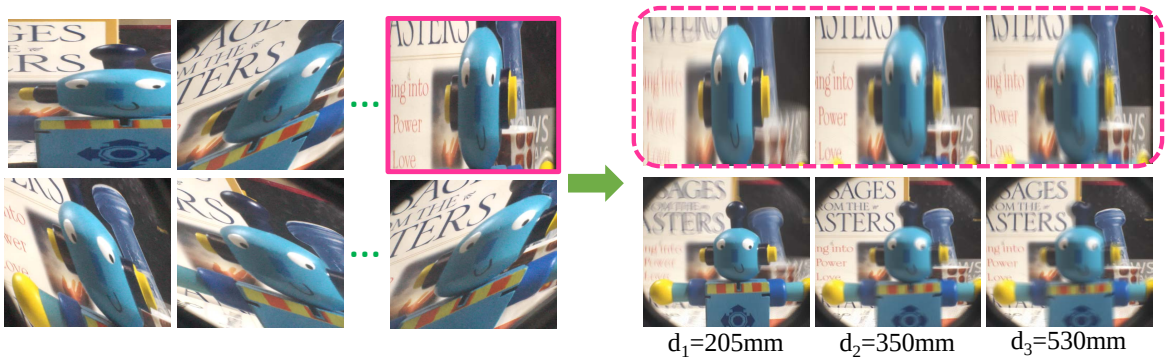
frustum matching the extent of both horizontal and vertical slits. To measure the color consistency, we need to first determine the projection of the voxel in each XSlit view. We use the XSlit projection Eqn. 3.14 to map every voxel to all XSlit cameras.

The voxel depth assignment problem is solved via the graph-cut algorithm [12, 11]. Specifically, we traverse spatial voxels through plane sweeping. For each voxel, we fetch corresponding pixels from respective XSlit images and compute their color variance as the data cost. We also adopt color weighted smooth priors for depth estimation. Fig. 3.10 shows the reconstruction results.

3.2.2 Scene Rendering

After we get the scene geometry, we can synthesize the new refocused images which are focusing on the objects in the scene. For 4D plenoptic sampling acquired by a pinhole camera array, the refocusing results are synthesized by interpolation of sampled images. This can be done by first imposing a geometry proxy, e.g., a 3D plane (as shown in the lumigraph [20]), then projecting rays from a reference view to intersect with the proxy, and finally tracing the intersections back to sampled images to fetch recorded radiances. Alternatively, one can use a disparity value, if epipolar geometry exists, to directly represent proxy geometry and to query corresponding pixels from the views. As discussed in Section 3.1.3, since there is no homogenous epipolar geometry in the R-XSlit sampling, we adopt the first scheme to render focus stacks.

XSlit Refocusing. For the R-XSlit sampling $\{\mathcal{C}_\theta | \theta \in \Omega_\theta = \{\beta_1, \beta_2 ... \beta_N\}\}$, we render the refocusing result $\mathcal{J}_\beta^f$ corresponding to the XSlit view $\mathcal{C}_\beta$, where superscript f indicates that the focal depth is $z_f = f$. Specifically, we first specify a geometry proxy plane and conduct backward tracing for view blending. Alternatively, we implement forward projection of each XSlit image onto the proxy plane and then combine all images via multi-texturing using the graphics pipeline. In fact, the forward projection of an XSlit image to an arbitrary 2D plane corresponds to a collineation that can be efficiently computed. We can further control the aperture size by changing the number of views involved in the blending.



(a) Sub-Xslit Images (b) Dynamic Refocusing

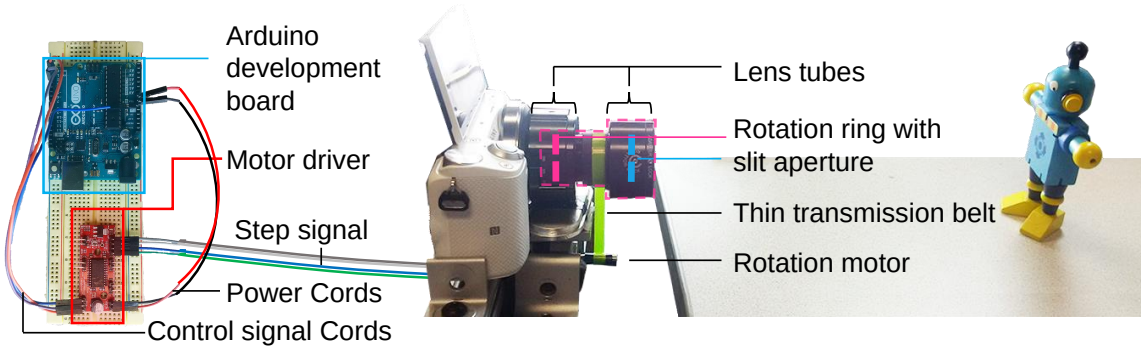
Figure 3.5: Dynamic refocusing images rendered from the R-XSlit sampling. (a) Sub-XSlit images are captured by our prototype R-XSlit camera. (b) Two different rendering effects. The first row shows the focus stack using a sub-XSlit image as a reference image; the second row shows refocusing rendering from a perspective view.

Using a small number of views will produce an image with deep depth of field whereas a large number will produce shallow depth of field.

Perspective Refocusing. With the R-XSlit sampling, we can also render a new perspective image focusing at a focal depth $z_f = f$. We sample a grid of voxels on the plane z_f to render a perspective image. For each voxel $X = (x, y, z_f)$, we trace the rays back to all the XSlit views to fetch recorded radiances. According to the projection Eqn. 3.14, we can compute the pixel location q_θ at $\mathcal{I}_\theta$ corresponding to X. Thus, the refocusing image $\mathcal{J}_P^f$ can be rendered as:

$$\mathcal{J}_P^f(p) = \frac{1}{N} \sum_{\theta \in \Omega_\theta} \mathcal{I}_\theta(q_\theta). \tag{3.15}$$

The most notable difference between perspective over the R-XSlit sampling is the defocus blur kernel. Fig. 3.5(b) shows examples of perspective view refocusing. In particular, refocused images exhibit a "conic blur" effect ,i.e., the blurriness is more severe near the boundary than near the center as shown in Fig. 3.5(b). In Fig. 3.5, we conduct real refocusing on a double-slit rotational sampling, from which we can see nice blurring due to dense angular sampling.



(a) Motor Control System

(b) System setup

Figure 3.6: The prototype of our rotational XSlit camera system. (a) The control circuit for the rotation motor. (b) System setup overview.

3.3 Experiments

We validate our R-XSlit sampling scheme on both synthetic and real scenes. In this section, we first talk about our acquisition devices and our camera structure. Next, we address the calibration problem of R-XSlit sampling and evaluate the practicability of our scheme. We also show rendering and stereo matching results with different sampling densities.

3.3.1 Camera Construction

Fig. 3.6 illustrates our prototype R-XSlit camera. We mount XSlit lenses on a commodity interchangeable lens camera (e.g.Sony NEX-5T), and align the two cylindrical lenses orthogonally with two lens tubes. Each tube contains a rotation ring which can help us control the rotation degree of each slit precisely.

In [65], R-XSlit pairs are acquired though rotating the XSlit camera. However, this method only works when the number of captured data is small. To form a valid plenoptic sampling, we need to capture large numbers of images as accurate as possible. Nevertheless, it is difficult to eliminate or even evaluate the slight bias of the rotation axis when rotating the camera, and the accumulation of small errors can lead to huge inaccuracy. To address this problem, we mount each slit to lens tube with a rotation ring which can rotate by 360 degrees freely without affecting the tube. Instead of rotating the camera, we rotate the lens tube. Moreover, to minimize the inaccuracy, we adopt a stepper motor to control the rotation

procedure. The lens tube and the motor lever are connected by a flat ribbon to make sure that they rotate in the same speed. To control the rev rate, we employ a Arduino Uno R3 board, i.e., a board that can control the rotation mode of stepper motor with an uploaded program from the computer. By applying the stepper motor to the XSlit camera, we are capable of capturing the R-XSlit plenoptic sampling through a video mode. Thus, we can minimize manual errors and sample rays without moving the camera.

Another advantage of adopting a stepper motor is that it is easy for us to control the density of the sampling. In our setup, we set the rotation rate at 12 degrees per second and the frame rate at 30. Typically, we can capture about 900 images for each plenoptic sampling when rotating the lens tube by 360 degrees.

To ensure the stability of the rotation, we adopt an additional calibration step beforehand. We capture 3 sets of R-XSlit images of a checkerboard calibration target, each with a different rotating speed of the motor. We then extract their corners for verification and find that the the views align almost perfectly with the theoretical computation. In fact, if they were not aligned due to the uneven rotation speed, the results could also be used to adjust the sequence in the following experiments. The wiggle of the axis also seems to have a slight effect on the results, which we suspect this is due to the rigidity of the camera and stepper motor which makes the jiggles nearly negligible. Finally, the lens tube sets are sealed to the camera body, and we have not observed obvious changing stray light patterns during the acquisition.

In terms of improving light accumulation, we adopt the dual cylindrical lens design and focus adjustment schemes [63] which have significantly improved the light throughput.

3.3.2 Calibration

Rather than trying to align the optical axis (i.e., the central ray), we set out to calibrate the camera by finding out the bias d_1, d_2 of l_1 and l_2. The two slits' position w.r.t.the image sensor are $Z_1 = 62mm$ and $Z_2 = 26mm$with a width of $2mm$. For a 3D point $X = (x, y, z)$ in a scene, we capture it three times by rotating the lens tube by 90 degrees on a rotation ring

(a) d_1=-0.07mm d_2=0.258mm (b) d_1=-0.17mm d_2=0.258mm (c) d_1=-0.07mm d_2=0.358mm

Figure 3.7: The refocusing images under different d_1, d_2. (a)(b) have 0.1 difference in d_1, (a)(c) have 0.1 difference in d_2.

to generate 3 XSlit images. According to Eqn. 3.13, the projection locations of X on image sensor should be:

$$\begin{aligned} u_0 &= \frac{Z_1 x - d_1 z}{Z_1 - z} \quad v_0 = \frac{Z_2 y - d_2 z}{Z_2 - z} \\ u_{90} &= \frac{Z_2 x + d_2 z}{Z_2 - z} \quad v_{90} = \frac{Z_1 y - d_1 z}{Z_1 - z} \\ u_{180} &= \frac{Z_1 x + d_1 z}{Z_1 - z} \quad v_{180} = \frac{Z_2 y + d_2 z}{Z_2 - z} \end{aligned} \tag{3.16}$$

By solving Eqn. 3.16 we can get:

$$\begin{aligned} d_1 &= -\frac{(u_{90} + v_0)(u_0 - u_{180})(Z_1 - Z_2)}{2Z_2(u_{90} - u_{180} + v_0 - v_{90})} \\ d_2 &= \frac{(u_0 + v_{90})(v_0 - v_{180})(Z_1 - Z_2)}{2Z_1(u_0 - u_{90} - v_{90} + v_{180})} \end{aligned} \tag{3.17}$$

We therefore choose 30 calibration points on $\mathcal{I}_0$ and find their corresponding points on $\mathcal{I}_{90}$ and $\mathcal{I}_{180}$ respectively. From Eqn. 3.16 and Eqn. 3.17, we derive 30 sets of (d_1, d_2). $d_1 = 0.05mm$, $d_2 = 0.28mm$ are the average values of the 30 results.

Fig. 3.7 illustrates that a slight bias of l_1 and l_2 will have a significant impact on the rendering performance. It is worth mentioning that the average value does not guarantee the optimal solution. To find out the correct (d_1, d_2), we first use the average d_1 and d_2 to generate a focus stack through Eqn. 3.15. Next, we pick out a slice that focuses on a highly textured object at the depth of f. Note that the slice might still be a little blurry due to incorrect d_1, d_2 values. We then crop 10 8x8 patches from the object, and use the focusness detection methods in [34] to measure the patches' focusness degree when varying d_1 and d_2 respectively. A focus degree for a pair (d_1, d_2) is calculated by averaging all the pixel values in the 10 focusness maps, and the (d_1, d_2) that achieves the highest degree is the optimal

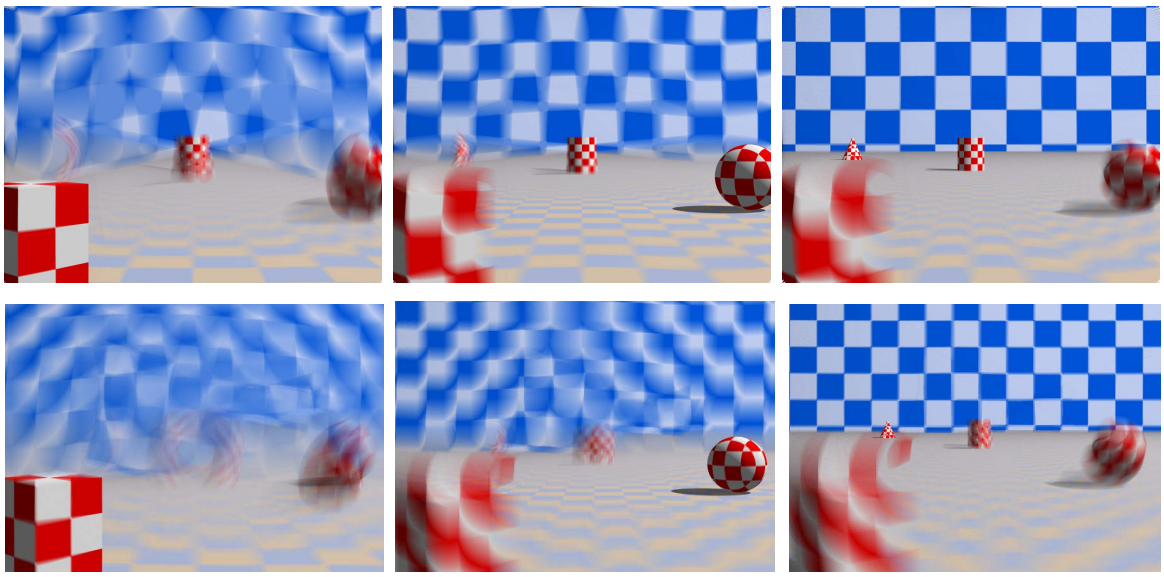


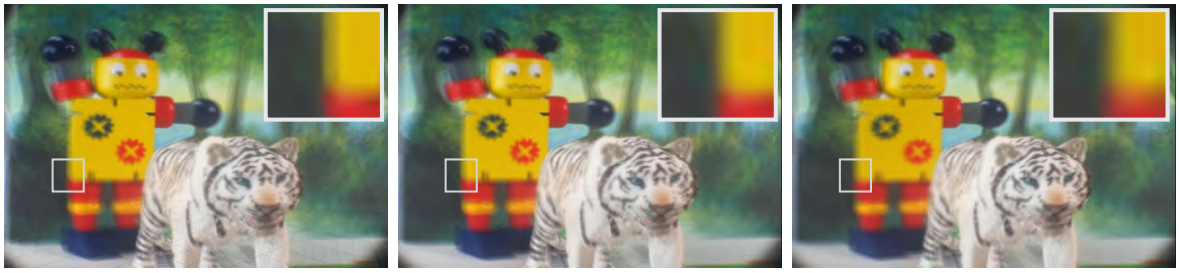
Figure 3.8: The refocusing effect using different R-XSlit camera settings. The first and second rows show the results corresponding to $\mathcal{C}_1$ and $\mathcal{C}_2$ respectively. (See text for details.)

solution. After the optimization procedure, we derive the best solution $d_1 = -0.07mm$, $d_2 = 0.26mm$.

3.3.3 Results

We conduct 3D reconstruction and refocusing rendering on both synthetic and real data.

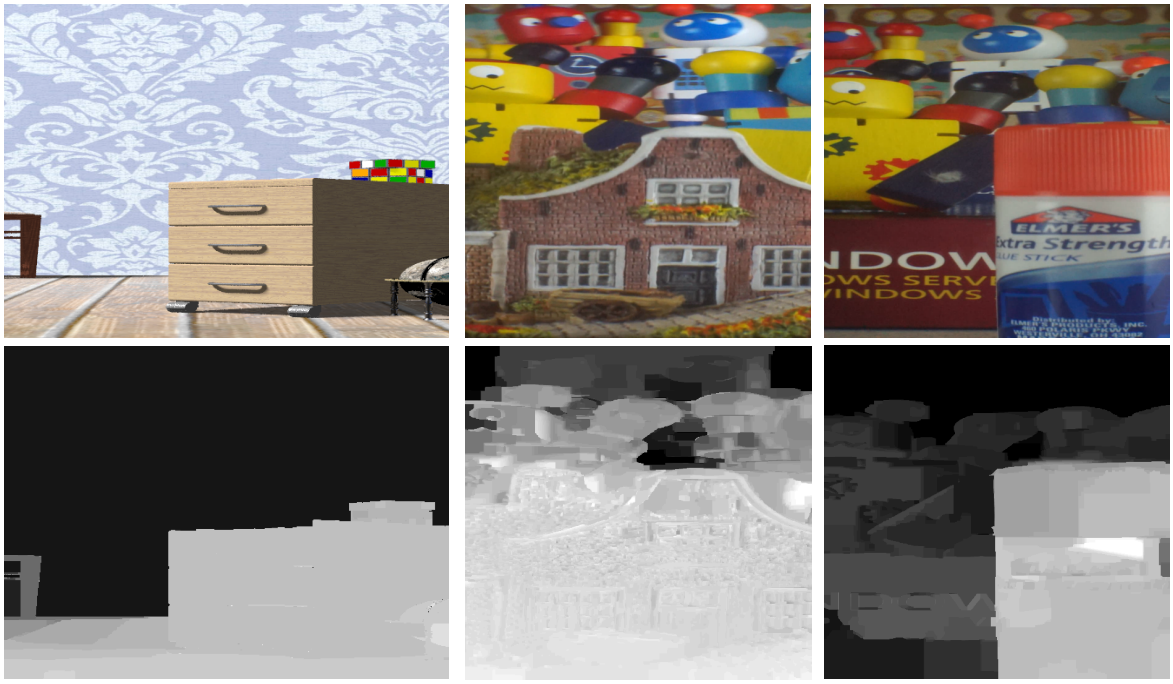
Synthetic Data. We first test our scheme on synthetic data rendered by the POV-Ray ray tracer. Fig. 3.8 presents the refocusing effects rendered by R-XSlit cameras $\mathcal{C}_1(-2, -6, \theta + 90^o, \theta, -0.2, 0.1)$ and $\mathcal{C}_2(-2, -6, \theta + 90^o, \theta, 0, 0)$. We collect 360 views by $\mathcal{C}_1$ and $\mathcal{C}_2$ with equal angular interval $\Delta\theta = 1$. In $\mathcal{C}_2$ case, $\mathcal{I}_\theta = \mathcal{I}_{\theta+180}$. For refocusing results from $\mathcal{C}_2$, the center portion is always in focus. This is because that when $d_1 = d_2 = 0$, the image centers of all sub-XSlit images correspond to a same ray. In contrast, $\mathcal{C}_1$ captures multiple rays for every pixel. The Conic Blur effect of $\mathcal{C}_2$ is more obvious than $\mathcal{C}_1$. It is worth noting that for the same reason, $\mathcal{C}_1$ achieves better reconstruction results than $\mathcal{C}_2$ for the center portion. Fig. 3.10 shows the depth reconstruction result of a synthetic example (in the first row) with $\mathcal{C}_2$.



(a) 50 sub-XSlit images (b) 200 sub-XSlit images (c) 500 sub-XSlit images

Figure 3.9: Refocusing rendering results using different sampling density along the rotation angle. In this example, we focus on the head of the tiger. The out-of-focus region is smooth even using a small number of sub-XSlit images.

Real Data. Next, we validate our model on scenes acquired by our R-XSlit prototype $\mathcal{C}_\theta(62, 26, \theta + 90^o, \theta, -0.07, 0.26)$ (Section 3.3.1). The R-XSlit sampling is captured through video recording. For each sampling capture, we can extract about 900 XSlit images at resolution 1920×1080 when two slits rotate 360^o. Fig. 3.9 presents the refocusing using different numbers of XSlit images and we can see that by incorporating more sub-XSlit images, some alias such as the black lines caused by insufficient sampling can be eliminated. However, the out-of-focus region is overall smooth even using a small number of sub-XSlit images. Fig. 3.10 shows the depth reconstruction results of some real scenes.



(a) (b) (c)

Figure 3.10: Depth reconstruction from the R-XSlit sampling on a synthetic example (a) and real examples (b)(c). The first row presents XSlit images and the second row shows their corresponding depth maps.

Chapter 4

3D RECONSTRUCTION FROM WAVELENGTH CODED PLENOPTIC SAMPLING

In this chapter, we present two wavelength coded plenoptic sampling schemes. We first exploit a concentric wavelength coded (CWC) plenoptic sampling scheme in visible spectrum. Our sampling system design employs concentric rings of cameras[1], with each ring capturing a narrowband spectrum, as shown in Fig. 4.1. We also use a single ring of lights, with each light at a different spectrum. We show this concentric setup imposes useful constraints on specularity variations that can be used to robustly separate diffuse components from specular ones.

To model surface reflectance with specularity, we employ the Phong dichromatic model and integrate it into our concentric wavelength coded plenoptic function, we further estimate surface normal and spectral reflectance map based on our sampling scheme. Specifically, we introduce a multi-spectral surface camera (MSS-Cam) by extending the classical surface camera (S-Cam) [68]: each ray sample in the MSS-Cam originates from the same 3D point but at a different angle and with a different spectrum. We propose a new appearance consistency metric across the views and a robust confidence metric to separate Lambertian points vs. non-Lambertian ones. For the Lambertian points, we apply a multi-spectral photometric stereo technique to recover its normal. For the non-Lambertian points, we remove the rays that correspond to specularity and use the rest for surface normal estimation. Comprehensive experiments show that our technique can achieve high accuracy and robustness in geometry and spectral reflectance recovery, to the benefit of a wide range of vision and graphics tasks.

---

[1] In our implementation, we move a single camera to emulate the rings of cameras

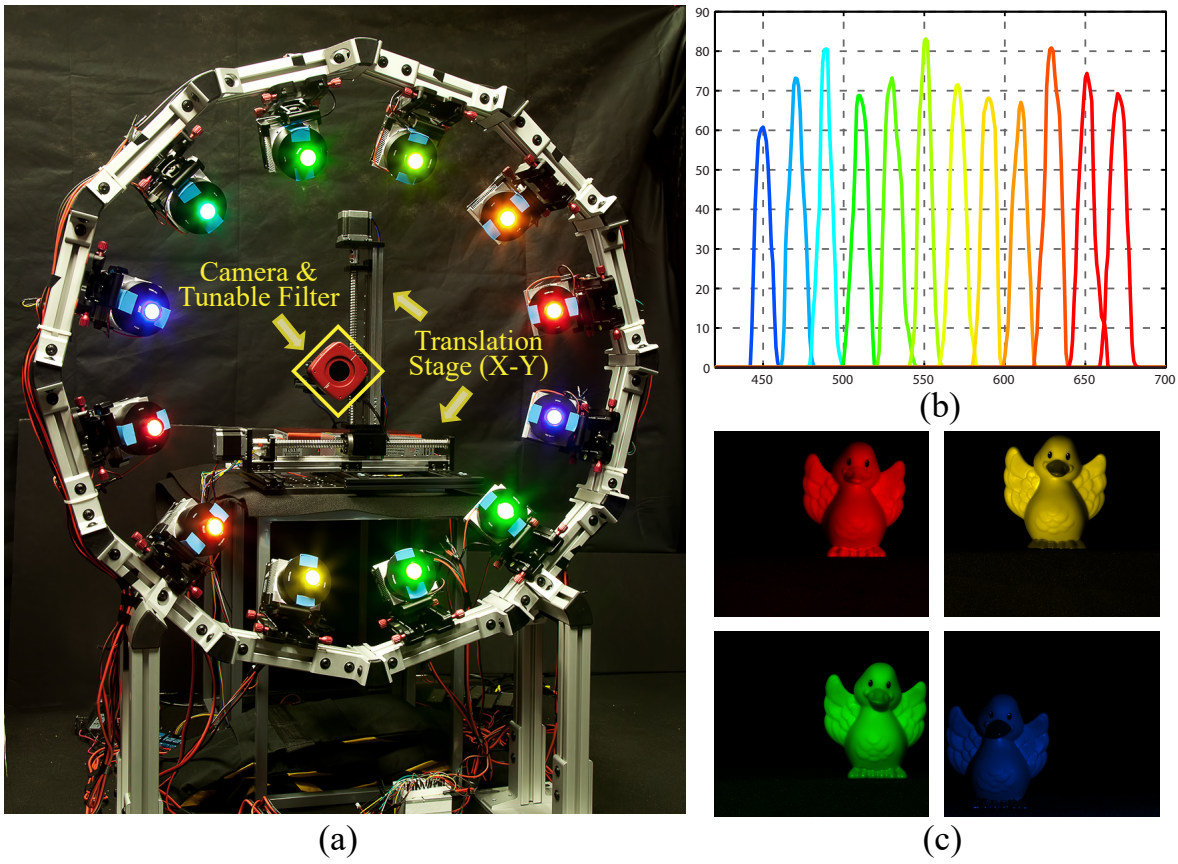


Figure 4.1: (a) Our CWC plenoptic sampling acquisition system; (b) The illumination spectral distribution; (c) Sample images from our sampling (We convert spectral images to RGB for better visualization).

Second, we present a infrared wavelength coded (IRWC) plenoptic sampling scheme, and we further propose a collaborative framework based on our sampling scheme to achieve pose estimation and face reconstruction under: (1) poor lighting conditions such as in complete darkness; (2) uncooperative conditions in which face images exhibit strong 3D pose orientations; (3) covert operations. Our reconstruction pipeline consists of 3 components: (1) 3D eye localization; (2) additional facial landmark detection; (3) 3D pose estimation and frontal face rendering. Our method is based on exploiting the "bright-eye" phenomenon that human eyes can be captured by NIR cameras with an NIR flash in the dark. The sampling system is consist of a pair of NIR cameras and a thermal camera where the NIR cameras are used to capture bright eyes and the thermal camera samples the long-wave infrared rays. The "bright" eyes are used to localize the 3D position of eyes and face. The thermal image provides additional facial points to address the 1D ambiguity in pose estimation. Consequently, the 3D pose orientations captured in the thermal face image are compensated in non-frontal face detection and recognition. Experiments on real face images are provided to demonstrate the merit of our method.

4.1 Object Reconstruction from CWC Plenoptic Sampling

4.1.1 Multi-Spectral Reflectance Model

To handle inhomogeneous surfaces with both diffuse and specular components, we adopt the Dichromatic Reflectance Model [50] (DRM) for material modeling. As DRM separates surface reflectance into body reflectance and interface reflectance and both terms account for geometry and color. DRM is suitable for modeling inhomogeneous materials.

Given a light source with the spectral distribution $E(\lambda)$ where λ refers to wavelength, and a camera with the spectral response function $Q(\lambda)$, the observed image intensity I under DRM at pixel p can be formulated as:

$$I(p) = w_d(p) \int_\Lambda R(p,\lambda)E(p,\lambda)Q(\lambda)d\lambda + w_s(p) \int_\Lambda E(p,\lambda)Q(\lambda)d\lambda \tag{4.1}$$

where $\Lambda = [\lambda_1, \lambda_N]$ is the range of sampled wavelengths; $R(p, \lambda)$ is the surface reflectance; $w_d(p)$ and $w_s(p)$ are geometry-related scale factors. The first term in Eqn. 4.1 represents body reflectance that models light reflection after interacting with the surface reflectance, and the second term represents interface reflectance that models light immediately reflected from the surface and thus causes specularites.

Eqn. 4.1 can be further discretized w.r.t. wavelength and written as:

$$I = w_d \mathbf{REQ} + w_s \mathbf{JEQ} \tag{4.2}$$

where $\mathbf{J}$ is a row vector with all ones,

$\mathbf{R} = [R(\lambda_1), R(\lambda_1 + \tilde{\lambda}), ..., R(\lambda_N)]$,

$\mathbf{E}(\mathbf{p}) = diag(E(\lambda_1), E(\lambda_1 + \tilde{\lambda}), ..., E(\lambda_N))$,

$\mathbf{Q} = [Q(\lambda_1), Q(\lambda_1 + \tilde{\lambda}), ..., Q(\lambda_N)]^T$.

Considering the scene geometry, we adopt the Phong dichromatic model that applies the classical Phong model on the top of the DRM (similar to [53]). Specifically, we model body and interface reflectance in terms of surface normal and roughness. Since body reflectance encodes the surface albedo while interface reflectance relates to specularity, we use the diffuse term as $w_d(p)$ and specular term as $w_s(p)$. With near point lighting (NPL), light source locations are used for computing lighting directions. Thus we rewrite the image intensity I as:

$$I = \alpha\Big(\frac{\hat{L} \cdot N}{\|L - X\|^2}\Big)\mathbf{REQ} + \beta\Big(\frac{(L_r \cdot V)^m}{\|L - X\|^2}\Big)\mathbf{JEQ} \tag{4.3}$$

where N is the surface normal at a 3D point X, L is the position of light source, $\hat{L} = (L - X)/\|L - X\|$ is the normalized lighting direction, V is the viewing vector, $L_r = 2(\hat{L} \cdot N)N - \hat{L}$ is the reflection direction, m is the shininess parameter that models the surface roughness, α and β correspond to the diffuse and specular reflectivity of the surface.

4.1.2 CWC Plenoptic Sampling

We then apply the above reflectance model to our CWC plenoptic sampling scheme and perform a specularity analysis under surface cameras (S-Cams) to distinguish diffuse

components vs. specular ones.

4.1.2.1 Sampling Scheme

We design a novel computational illumination/imaging system for plenoptic sampling, as shown in Fig. 4.2.

Our design mounts concentric rings of cameras, with each ring capturing a narrowband spectrum. In particular, the cameras lie on m concentric circles with radiuses $r_j|j = 1, ..., m$. On each circle, there are n cameras arranged with the angle interval $\tilde{\phi}$. We assume that all CoPs of the cameras lie on a common plane ($z = 0$). In order to capture the irradiances coming from different spectra, each camera is mounted with a spectral filter. The cameras on the same ring have the same spectral filter with wavelength λ_j, where j is the concentric circle index. The set of viewing directions for cameras on the jth circle is $\mathbf{V_j} = [V_1^{(j)}, ..., V_n^{(j)}]^T$. We will later show in our specular analysis that such concentric camera arrangement benefits the separation of diffuse components and specular ones.

We adopt the two-plane parameterization (2PP) [33] where rays are parameterized by their intersections with two parallel planes. We assume the st plane is at $z = 0$ and the uv plane is at $z = 1$, we fix the uv plane on the image plane of cameras (assume that all cameras have the same focal length that is normalized to 1). We use $(s, t) = (r_j \cos\phi_i, r_j \sin\phi_i)$, where $\phi_i = \phi_1 + (i-1)\tilde{\phi}$, to index camera position in the sampling and employ (u, v) as pixel index in each captured image. Therefore, our CWC plenoptic function can be formulated as follow:

$$I = P(u, v, r_j, \phi_i, \lambda_j) \tag{4.4}$$

In addition, we use a single ring of lights, with each light having a different spectrum, to provide multi-spectral illumination. In particular, we have m point light sources located on a circle with a radius of r_l, and the angular interval is $\tilde{\theta}$ with respect to the center of the circle. Notice that the number of light sources is the same as that of the camera concentric circles. Further, our lighting spectra have wavelength $\lambda_j|j = 1, ..., m$, which is in the same spectral space as the camera. So each spectral light corresponds to a ring of sampling cameras.

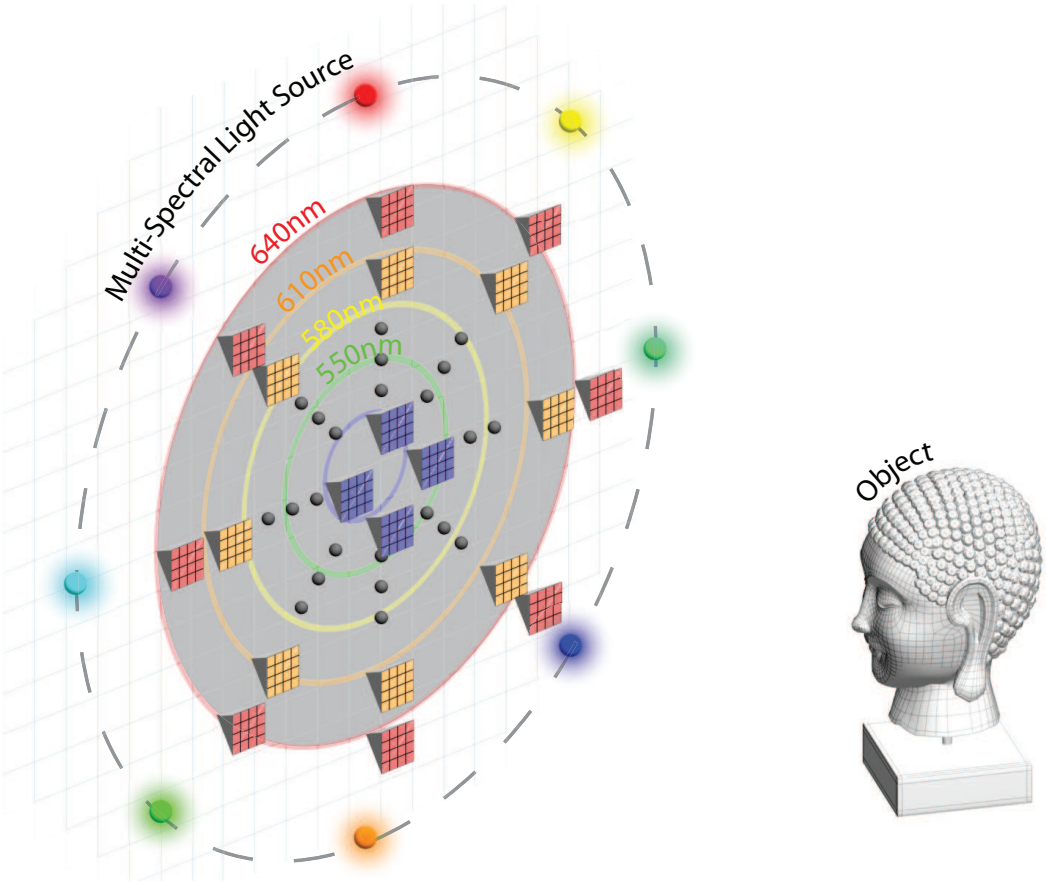


Figure 4.2: The CWC plenoptic sampling scheme.

Next, we derive the lighting directions that are critical to the reflectance model. In the angular domain, we represent the light source location as $\theta_j = \theta_1 + (j-1)\tilde{\theta}$ where θ_1 is the angular position of the first light source. Assume that these light sources are also placed at the plane $z = 0$, the set of light positions is $\mathbf{L} = [L_1, ..., L_m]^T$ where $L_j = [r_l \cos\theta_j, r_l \sin\theta_j, 0]$, and the normalized lighting direction set is $\hat{\mathbf{L}} = [\hat{L}_1, ..., \hat{L}_m]^T$.

4.1.2.2 Multi-spectral Surface Camera (MSS-Cam)

Next, we apply the NPL Phong dichromatic reflectance model derived in Sec. 4.1.1 to our CWC plenoptic sampling and perform specular analysis under the S-Cam [68]. The Surface Camera or S-Cam [69] characterizes the angular sampling characteristics of a LF from a 3D scene point. Given a point in a 3D scene, its S-Cam can be synthesized by tracing rays originated from the scene point into the LF to fetch color.

Applying the S-Cam to our plenoptic sampling, we obtain the multi-spectral S-Cam or MSS-Cam. We now derive intensities captured by the MSS-Cam using our reflectance model. Given a pixel in the reference center view with the camera position $(s,t) = (0,0)$, assume that its corresponding 3D scene point is $X = (x, y, z)$, we can synthesize its MSS-Cam M_X from our plenoptic sampling. As shown in the Fig. 4.3 (Left), each column of M_X is sampled under the same spectrum due to our unique camera/light source arrangement. The column therefore samples specularity variations from a common circle and exhibits periodical changes in intensity. To obtain M_X, we trace rays from the point X to each camera in the CWC plenoptic sampling. For a pixel (i,j) in our MSS-Cam, its sampling ray is from the camera at (s_i, t_j), the projection of X on the image I_{ij} is $p = (u_{ij}, v_{ij})$ and the intensity can be computed by bilinear interpolation. Therefore, with Eqn. 4.3 and Eqn. 5.12 we have the image intensity at $M_X(i,j)$ as:

$$\begin{aligned} M_X(i,j) &= P(u_{ij}, v_{ij}, r_j, \phi_i, \lambda_j) \\ &= \alpha_X \Big(\frac{\hat{L}_j \cdot N_X}{\|L_j - X\|^2} \Big) c_X \mathbf{B_j E_j Q_j} \quad + \beta_X \Big(\frac{(L_r^{(j)} \cdot V_{i,j})^{m_X}}{\|L_j - X\|^2} \Big) \mathbf{J E_j Q_j} \end{aligned} \tag{4.5}$$

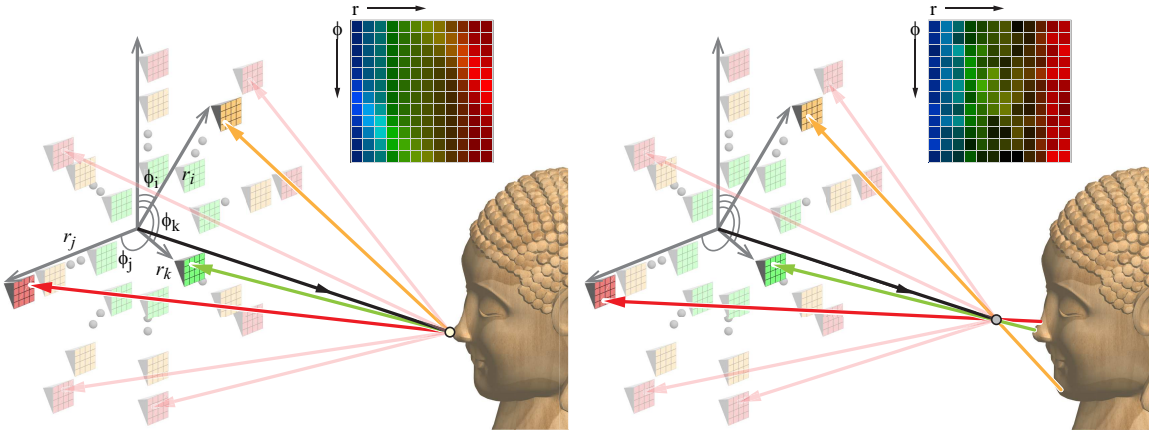


Figure 4.3: MSS-Cam. Left: the example of our MSS-Cam with a correct depth. Right: the MSS-Cam for the same point with an incorrect depth.

where N_X and m_X are surface normal and roughness of the X; the reflectance spectra R in Eqn. 4.3 can lie in a w-dimensional linear subspace [46, 36], thus the $c_X = [c_1, ..., c_w]$ denotes the reflectance coefficient vector; B_j is a $w \times k$ linear reflectance basis matrix under spectral range $[\lambda_j - (k-1)\tilde{\lambda}/2, \lambda_j + (k-1)\tilde{\lambda}/2]$. E_j and Q_j are also under this spectral range with a size of $k \times k$ and $k \times 1$ respectively. We combine the sampling rays from all cameras to form the M_X of the MSS-Cam at the scene point X.

4.1.2.3 Diffuse vs. Specular Analysis

We perform a photo-consistency analysis on the MSS-Cam and develop a robust confidence metric for separating diffuse points and specular ones.

We define a consistency measurement on the MSS-Cam using the standard deviation of intensities:

$$C(M_X) = \frac{1}{m} \sum_{j=1}^{m} std(M_X(1, j), ..., M_X(n, j)) \tag{4.6}$$

where $std(\cdot)$ is a standard deviation function. For Lambertian points, C is close to 0 when depth is correctly estimated. Therefore, we apply the Peak Ratio analysis [25] and use a certain threshold t_1 to separate the Lambertian points.

The points with a large first consistency measurement may be non-Lambertian, occluded and shadowed points. We introduce a second consistency measurement to refine the

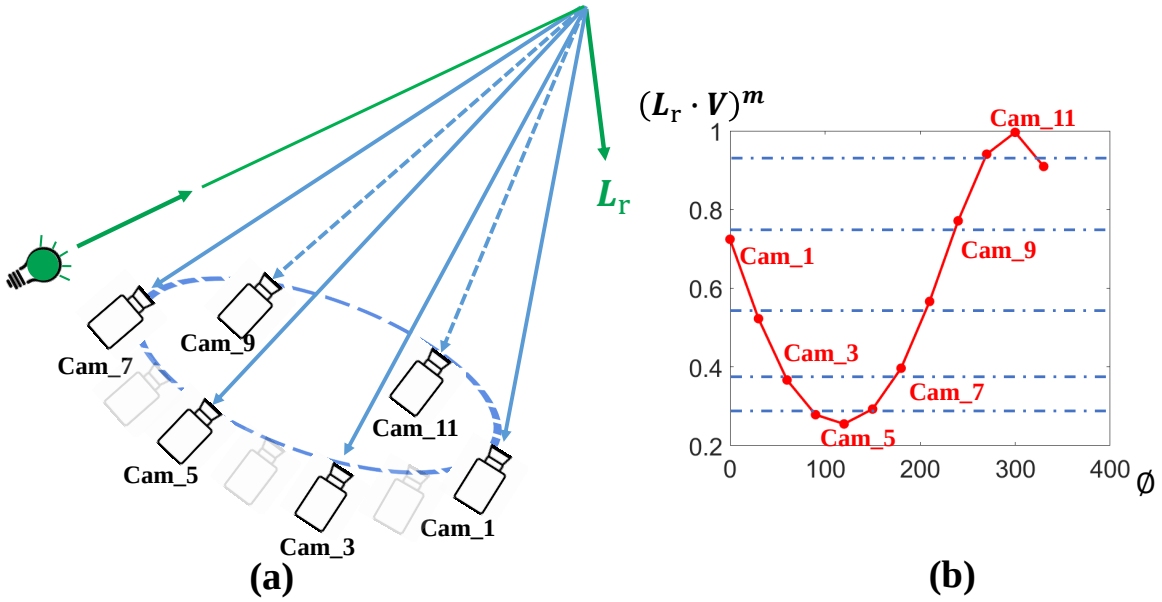


Figure 4.4: The periodical property of specularity. (a) Light sources configuration w.r.t. a scene point; (b) Measured sepcular components from views are on a periodic curve.

depths of non-Lambertian points. Since the cameras sampling for each spectrum are on a circle, the specular components will change on a periodic curve as shown in Fig.4.4. Therefore, the curve should be symmetric at its peak or valley. Recall that the specularity variation under one spectral light source is recorded in a column in our MSS-Cam. We extract the positions of the first three maximums as order p_{ml}, p_m and p_{mr}, where p_{ml} and p_{mr} are the left and right positions to the maximum, the second consistency measurement can be defined

as:

$$
\begin{aligned}
p_l &= \begin{cases} p_{ml}, & \text{if } |p_{ml} - p_m| < |p_{ml} - p_{mr}|. \\ p_m, & \text{otherwise.} \end{cases} \\
p_r &= \begin{cases} p_{mr}, & \text{if } |p_{mr} - p_m| < |p_{ml} - p_{mr}|. \\ p_m, & \text{otherwise.} \end{cases} \\
M_d(a,j) &= M_X(p_l - a, j) - M_X(p_r + a, j) \\
D(j, M_X) &= \sum_{a=0}^{(n/2-1)} (M_d(a,j) - M_d(a+1,j))
\end{aligned}
\tag{4.7}
$$

Note that the second consistency measurement on each column of the MSS-Cam should be small for specular points and large for some occluded or shadowed points. Thus, by simply applying a certain threshold t_s, we can filter out the occluded or shadowed points from specular points. We can robustly obtain the set of diffuse points and specular points by combining the two consistency measurements.

Specular Removal. For non-Lambertian points, we exploit specular variations across sampling views to remove specularity. Specifically, given the pre-calibrated term $\mathbf{JEQ}$ and an estimated depth z' for X, we compute the vertical gradients of the MSS-Cam to remove diffuse components in Eqn. 4.5 as:

$$
\begin{aligned}
G_X(i,j) &= \Big(M_X(i+1,j) - M_X(i,j)\Big) \frac{\|L_j - X\|^2}{\mathbf{JE_jQ_j}} \\
&= \beta_X((L_r^j \cdot V_{i+1,j})^{m_X} - (L_r^j \cdot V_{i,j})^{m_X})
\end{aligned}
\tag{4.8}
$$

We use the observation $\tilde{G}$ as the specular constraint to optimize surface normal, specular reflectivity and surface roughness simultaneously by:

$$
\underset{N_X, m_X, \beta_X}{\operatorname{argmin}} \sum_{i,j} \|\tilde{G}_X(i,j) - \beta_X((L_r^j \cdot V_{i+1,j})^{m_X} \qquad - (L_r^j \cdot V_{i,j})^{m_X})\| \tag{4.9}
$$

Since this nonlinear optimization is very complex, we try to remove specularities from our MSS-Cam instead of solving all variables simultaneously. In particular, we first

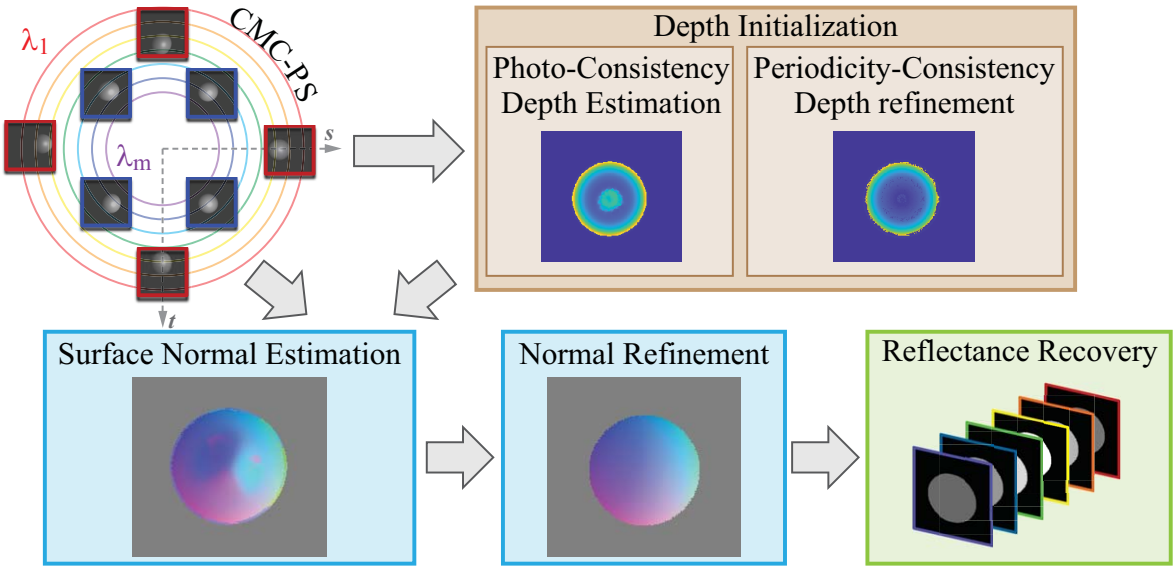


Figure 4.5: Our shape and reflectance reconstruction pipeline

remove points that are highly specular, and then solve parameters for specular-free points with the standard Levenberg-Marquardt method. The solver forces the parameters to fit the periodical specularity variation curve.

4.1.3 Shape and Reflectance Reconstruction

Finally, we use our specular analysis for surface shape and reflectance reconstruction. Our reconstruction pipeline is shown in Fig. 4.13.

Given an point X and its pixel p in the reference center view, we first compute our proposed photo-consistency measure C for every hypothetical depth z of X, and then initialize the depth z of X corresponding to the lowest measurement:

$$z' = \underset{z}{\operatorname{argmin}}\, C(M_X^{(z)}) \tag{4.10}$$

Then, we classify this pixel as Lambertian or non-Lambertian point via the consistency value. If it is classified into the non-Lambertian points, we apply our periodicity consistency to refine its depth as:

$$z' = \underset{z}{\operatorname{argmin}} \frac{1}{m} \sum_{j=1}^{m} D(j, M_X^{(z)}) \tag{4.11}$$

For any non-Lambertian point with an estimated depth, we retrieve its MSS-Cam and remove its specular components through optimizing its specularity variation. We then obtain specular-removal MSS-Cam, We use a multi-spectral photometric stereo method to recover surface normal and spectral reflectance coefficients as:

$$\underset{N_X, c_X}{\operatorname{argmin}}((c_X \mathbf{S}) \circ (\hat{\mathbf{L}} N_X)^T - \mathbf{M}) \tag{4.12}$$

where

$$c_X = [c_1, ..., c_w]$$
$$N_X = [n_x, n_y, n_z]^T$$
$$\mathbf{S} = [\mathbf{S_1}^T, ..., \mathbf{S_m}^T]$$
$$\mathbf{S_j} = \mathbf{B_j E_j Q_j} \quad , \quad \mathbf{S_j} = [s_{j1}, ...s_{jw}]$$
$$\hat{\mathbf{L}} = [\hat{L}_1, ..., \hat{L}_m]^T, \hat{L}_j = [l_{j1}, l_{j2}, l_{j3}]^T$$
$$\mathbf{M} = [M_X(p_1, 1), ..., M_X(p_m, m)]^T$$

p_j is the median position of the jth column on the MSS-Cam M_X and $\circ$ is the Hadamard product or the element-wise multiplication. When $3 \times w \leq m$, this bilinear optimization can

be formed into an over-determined linear least-squares optimization as:

$$
\begin{aligned}
A &= \begin{bmatrix} s_{11}l_{11} & s_{11}l_{12}\ldots s_{1w}l_{12} & s_{1w}l_{13} \\ s_{21}l_{21} & s_{21}l_{22}\ldots s_{2w}l_{22} & s_{2w}l_{23} \\ \vdots & & \\ s_{m1}l_{m1} & s_{m1}l_{m2}\ldots s_{mw}l_{m2} & s_{mw}l_{m3} \end{bmatrix} \\
b &= [c_1 n_x, c_1 n_y, c_1 n_z, c_2 n_x, \ldots c_w n_z]^T \\
b &= A \setminus \mathbf{M}
\end{aligned}
\tag{4.13}
$$

Surface normal and reflectance coefficients can be derived from b.

When $w \times 3 > m$, the linear least squares optimization can be transferred to over-determined bilinear optimization. Hence, we apply the general Levenberg-Marquardt algorithm to solve it.

To reduce the ambiguity caused by separating albedo and shading variation, i.e., the ambiguity on separation of the multiplication of two smooth curves, we arrange spectral light sources in the way in Fig. 4.6(a). Such arrangement can generate a fluctuated shading curve for the shading variation, shown in Fig. 4.6(b). Therefore, the surface normal can be more easily converged to the global optimum. We choose the median value of each spectrum on the MSS-Cam as the observation for optimization, which makes it more robust for image noise.

After we obtain all recovered surface normals for all pixels in the reference center view, we update the depths from the estimated surface normals. Finally, we renew all MSS-Cams with updated depths and use them to refine the surface normal and perform an iterative optimization until convergence:

Given the recovered estimated reflectance coefficients, we can recover the spectral reflectance of the object. Since the matrix $\mathbf{S_j} = \mathbf{B_j E_j Q_j}$ is calibrated from the spectral reflectance coefficients, and the matrix $E_j Q_j$ is simply regarding a single narrow-band spectral light source, we can directly use recovered reflectance coefficients c'_X to get dense spectral reflectance response by $\mathbf{R} = c'_X \mathbf{B}$, where $\mathbf{B}$ is the dense spectral sampling reflectance basis.

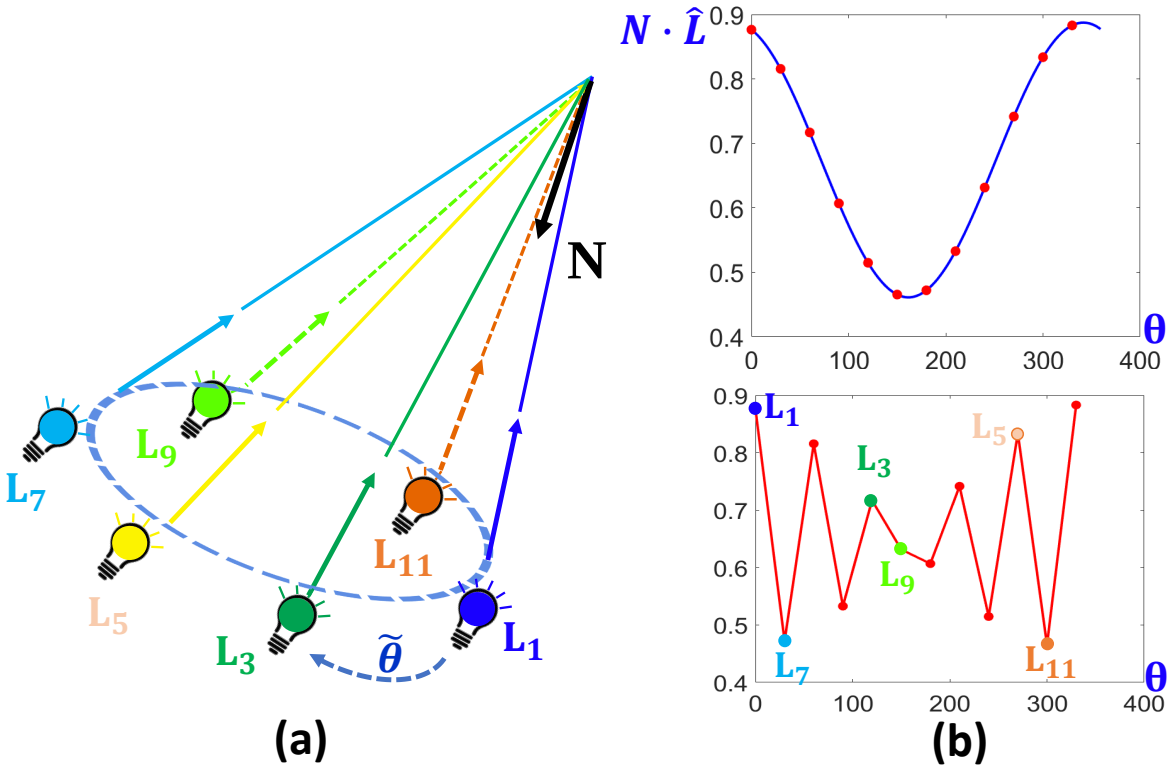


Figure 4.6: (a) Spatial relationship between a surface normal and all lighting directions; (b) The first row shows that values of shading components with the light arrangement are on a periodic curve. Therefore, we rearrange light positions to generate a fluctuated shading variation (second row) for robust separation between shading and reflectance.

4.1.4 Experiments

We have validated our approach on both synthetic and real data. All experiments are conducted on a desktop with an Intel i7 7820 CPU (2.9GHz Quad-core) and 32G memory. Our algorithm is implemented in Matlab, and all multi-spectral images are illustrated in RGB for better visualization.

4.1.4.1 Camera System Construction

We construct a multi-spectral camera array to evaluate our algorithm on real-world data. To build the multi-spectral camera array, we mount a monochrome camera (Point Grey GS3-U3-51S5M-C) with 50mm lens on a translation stage to uniformly translate the camera position on a 2D plane i.e., the st plane. We mount a tunable liquid crystal spectral filter

(KURIOS-WL1) in front of the camera to capture the scene under specified wavelengths. The camera resolution is 2448 $\times$ 2048 with a 13-degree FoV. Our hardware setup is shown in Fig. 4.1. To build the illumination part, we mount twelve 30Watt LED chips onto a circle dodecagon frame, and the distance between each LED chip and the center of the dodecagon is 90cm. We then place twelve narrow-band spectral filters ranging from 450 nm to 670 nm with 20 nm step in front of the LED chips. The distance between the acquisition system and the object is about 100 cm.

4.1.4.2 Calibration

We calibrate the position and radiance of each light beforehand, and the reflectance basis function is pre-computed during the calibration.

Camera Calibration. We first calibrate the camera intrinsic parameters using traditional camera calibration method. Our translation stage can provide high precision in translation and fix the orientation of the camera. The absolute value of the translation can be obtained by pre-captured a scene with a normal checkerboard.

Light Source Calibration. To calibrate the light position, we first move the camera to the center of the concentric circle. For each individual light, We need to capture a sequence of images with a chrome ball at the different positions. And in each image, we detect the specular spot on the chrome ball as the point of the incident. The corresponding incident ray can be derived from the reflected ray and its normal. Therefore, the position of the point light source can be localized at the intersection of all incident rays from those images. We need to repeat this procedure for all the lights. In reality, we capture the image with all lights on, and separate the light by their wavelength, so we can calibrate all the light positions at once.

Spectral Calibration. There are two parts for the spectral calibration. In the first place, we calibrate the $\mathbf{S} = \mathbf{BEQ}$ for each spectral light source by using a MacBeth ColorChecker chart. There are 24 diffuse color swatches on the chart with the ground truth spectral reflectance responses. We apply PCA to extract the w-dimension linear reflectance basis and the corresponding reflectance coefficient vectors $C_h = [c_{h1}, ..., c_{hw}]^T$

where $h = 1, ..., 24$. For each spectral light source, we capture the color checker at different orientations. With the known spectral light position, checker position and checker surface normal, we can get rid of the shading term $(\hat{L} \cdot N)/\|L - X\|^2$ to obtain the intensity I'. Therefore, the S_j for the jth spectral light can be attained by:

$$\underset{S_j}{\operatorname{argmin}}(\mathbf{C}S_j - I'_j) \tag{4.14}$$

where $\mathbf{C} = [C_1, ..., C_{24}]^T$ is a $24 \times w$ coefficients matrix, I'_j is the vector of the averaged intensities without shading components of all color patches under the jth spectral light source.

We then can obtain the set $\mathbf{S} = [S_1, ..., S_m]$. Note that we assume the diffuse reflectivity α of the color chart is 1 in the calibration, any other object's diffuse reflectivity will be assembled into the reflectance coefficient vector in the optimization.

In the next step, since we apply numerical integration to approximate the irradiance integration instead of using a Dirac Delta function, the integration of multiplication of the illumination spectral distribution and the camera response in the specular term can be attained from the calibration of the $\mathbf{E_j Q_j}$. In our experiments, the Full Width-Half Maximum (FWHM) of the filter is $10nm$, so we choose the step $\tilde{\lambda} = 1$ in the interval $[\lambda_j - 5, \lambda_j + 5]$ for integration. Therefore, the $\mathbf{E_j Q_j}$ can be obtained with a known 24×11 matrix $R_j = \mathbf{C}B_j$ as:

$$\underset{(\mathbf{E_j Q_j})}{\operatorname{argmin}}(R_j(\mathbf{E_j Q_j}) - I'_j) \tag{4.15}$$

The value of $\mathbf{J E_j Q_j}$ in the specular term for each spectral light source is known after this calibration.

4.1.4.3 Synthetic Results

We generate a multi-spectral renderer to render the multi-spectral images from RGB images based on a 3D reflectance linear basis and use spectral measures of real spectral light sources as the spectral distributions of our synthetic spectral light sources.

We first test our scheme on a simple sphere with three different reflectances, i.e., uniform diffuse, specular and specular with texture. The diffuse coefficients are all set to 0.7, and specular coefficients for each material are set to 0, 0.3, and 0.5 respectively. The

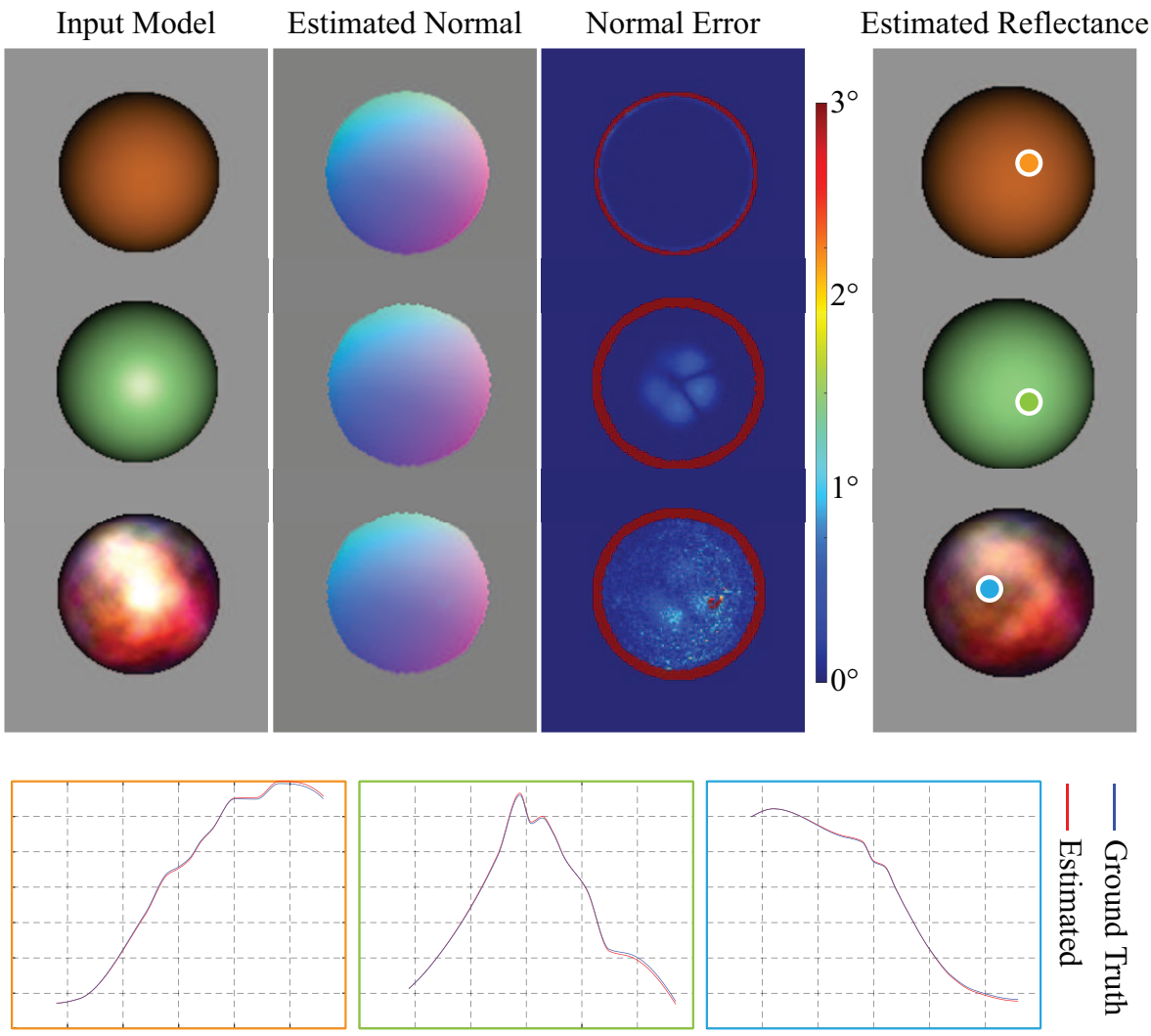


Figure 4.7: Qualitative synthetic results. The first column shows the input sphere with different reflectances. The second and third columns are our estimated normal maps and corresponding error maps. The last column is the estimated reflectance displaying in RGB, the dense recovered spectral reflectances compared to ground truth curves are presented at bottom.

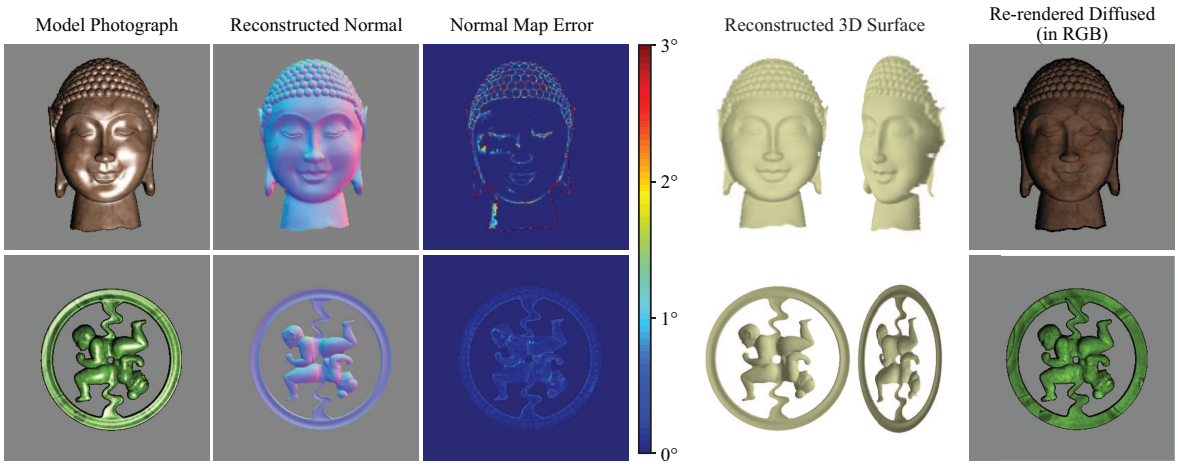


Figure 4.8: Shape and reflectance estimation on two complex synthetic scenes. The normal error maps and promising re-rendered diffused results demonstrate that our algorithm is robust against the specularity.

roughness for specular spheres is set to 8. In our experiment, we set the radius of the sphere to 20 and the distance between the cameras and sphere to 120. The resolution of each camera is 320×320. The radius of our light sources is 80, consisting of 12 spectral light sources. We capture a 12×12 CWC sampling, with the radius of the sampling circle ranging from 29 to 40 with step 1. Then, we set the synthetic wavelength of the filter between $440\ nm$ and $660\ nm$ with an interval of $20\ nm$ for both cameras and light sources. For depth estimation, we set the depth range between 108 and 125 with step 0.2, so that the sphere is modeled after 76 depth layers. We employ our consistency measurement to initialize the depth, which takes about 5 minutes, and then reconstruct surface normal and reflectance. Fig 4.7 shows the surface normal error map, we can see that all the degree errors are kept within 2 degrees. The bottom row in the Fig. 4.7 demonstrates examples of spectral reflectance estimation.

Next, we test on two more complex scenes. Since these two models have more complicate geometric structures, we also render the 12×12 CWC sampling with a higher resolution (500×500). Specifically, we set the specular coefficient to 0.6 and 0.4 respectively, and the roughness to 5. The wavelengths of filters are the same as those on the first scene. The distances between objects and cameras are 41 and 34. The radius of the CWC sampling ranges from 4 to 2.9 with step 0.1. The depths ranges are from 35 to 42 with step 0.1, and

from 33 to 35 with step 0.1 respectively. Fig 4.8 shows our reconstruction results. Obviously, our algorithm can achieve reasonable spectral reflectance and 3D recovered geometries, with the maximum normal errors kept within 3 degrees. The artifacts around the right eye of the Buddha head are caused by extra-low spectral reflectance response over all wavelengths.

4.1.4.4 Real Results

We first test our system for a simple plane scene with a color checkerboard, and the scene setup is shown in Fig 4.9. We extract the ground truth normal of the color checkerboard from its four corners. After reconstruction, we compute angular errors between ground truth normal and estimated normal of each color checker. The angular errors are small (less than 8 degrees), as shown on the upper right of Fig. 4.9. The white and black patches have the highest errors due to saturation and underexposure. To solve this problem, high dynamic range could help on.

We then test our reconstruction algorithm on four objects with different materials (ceramic, clay, plastic etc.). The results are shown in the Fig. 4.10 and Fig. 4.11. In the duck and sheep scenes, our results preserve high-frequency details of the sheep body and the duck's wing. Besides, the girl's face reconstruction from the last experimental result demonstrates that our algorithm is able to recover clean reflectance and surface normal for specular regions. We also test our algorithm on more challenging scenario, such as, the pure red in the strawberry scene and the pure green in the cactus scene, as shown in Fig. 4.11. For those scenes, our approach can still provide reliable reconstruction results.

For shadowed and occluded regions, since our spectral light sources are located on a circle, a shadowed 3D point can always be lit up by some parts of the light sources. We drop some spectral columns with large variations from our MSS-Cams before surface normal reconstruction and choose $w = 5$ and $m = 12$ for real scenes. Thus, we can drop at most 5 pieces of spectral information. However, for larger shadowed and occluded regions, we do not have enough spectral information to recover reflectance and surface shape simultaneously, such as the boundary around the dress in the last scene in the Fig 4.10 which is largely shadowed and occluded by the dress.

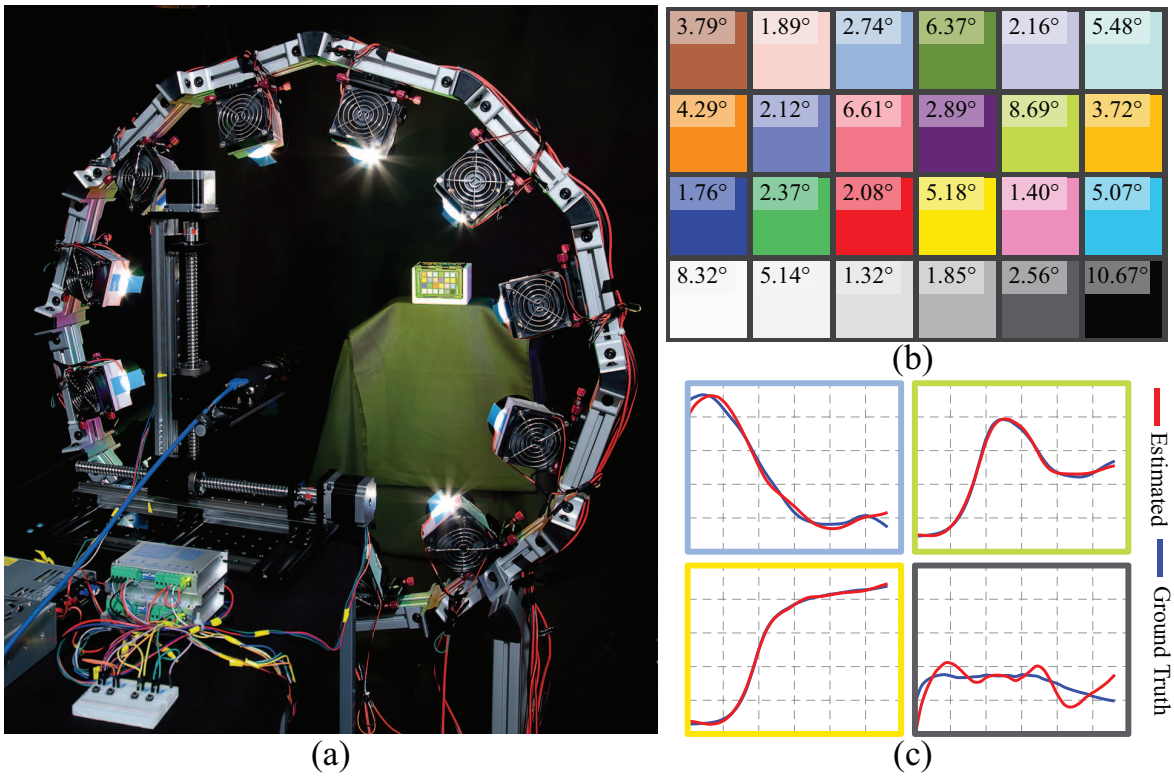


Figure 4.9: (a) scene setup. (b) the angular differences between estimated normal and ground truth normal in each color checker. (c) the estimated reflectance curves (in red) compared with the ground truth (in blue).

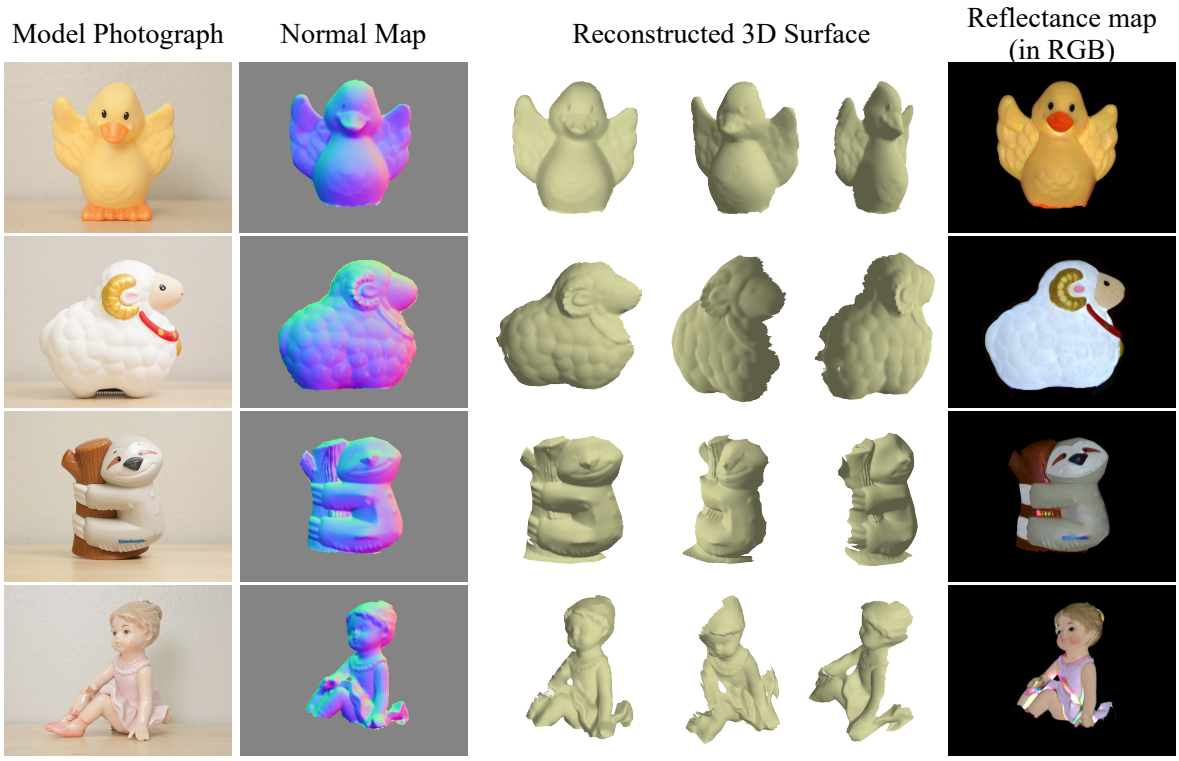


Figure 4.10: Shape and reflectance estimation results on real scenes with different materials. The first column is used to visualize models in RGB. The reconstructed shapes are in the third column. In order to visualize the recovered reflectance, we transfer the dense spectral reflectance to RGB relfectance, as shown in the last column. It can be seen that our approach can achieve favorable results.

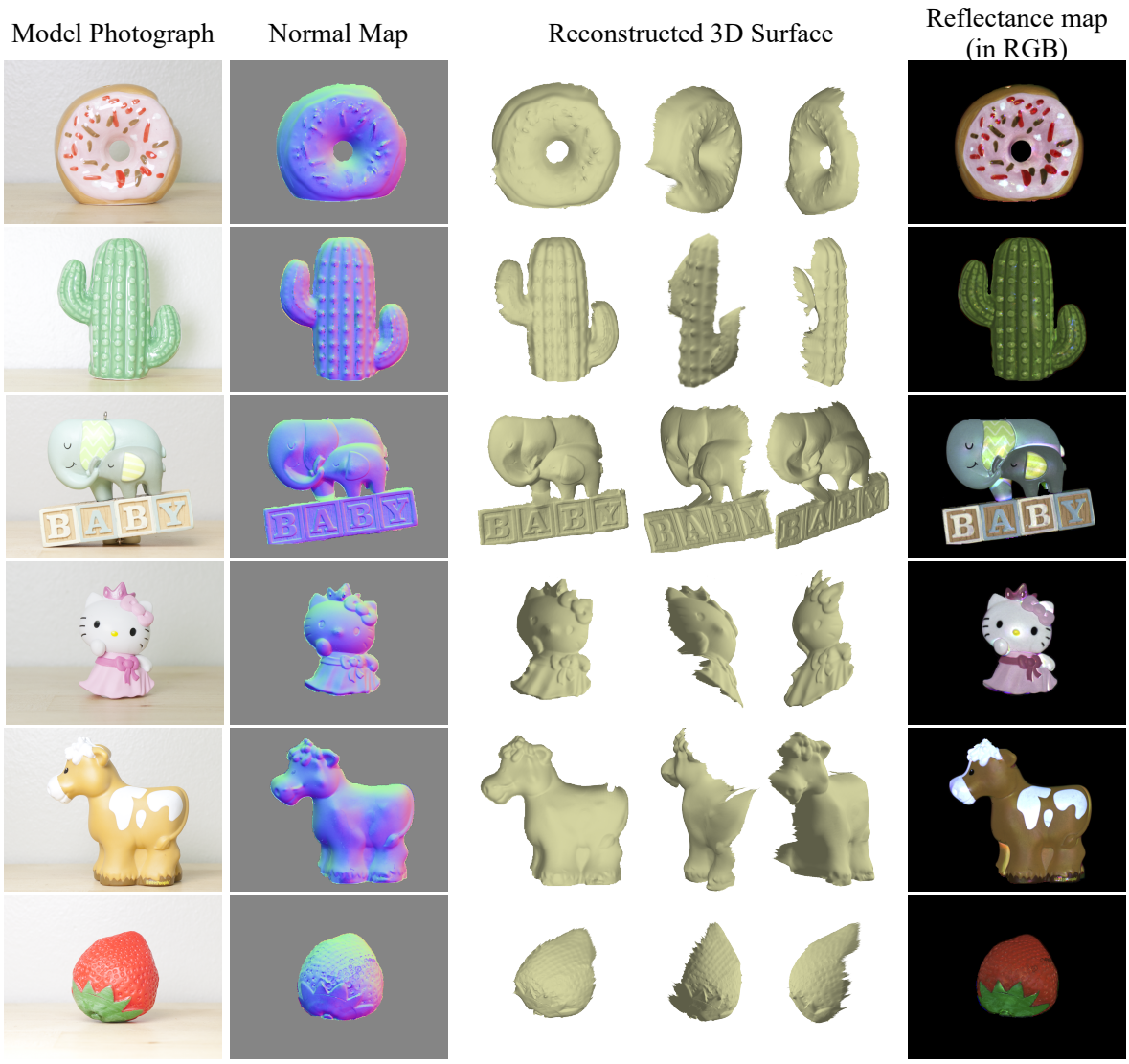


Figure 4.11: Additional shape and reflectance estimation results on real scenes with different materials. The reconstructed shape are represented at the third column and the RGB reflectances are shown in the last column.

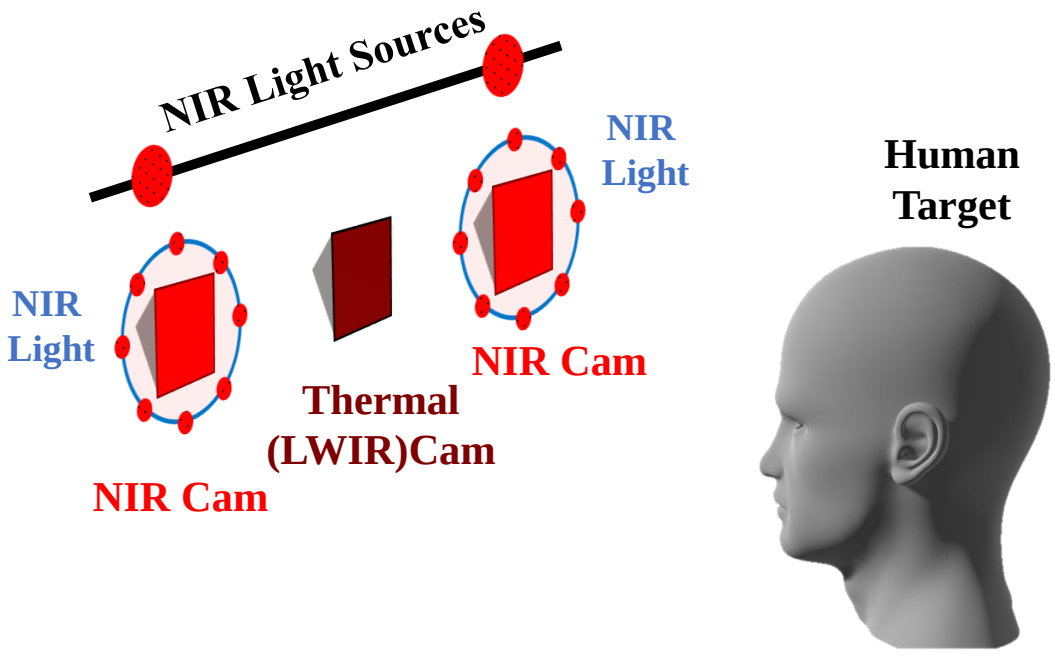


Figure 4.12: The IRWC plenoptic sampling scheme.

4.2 Face Reconstruction from IRWC Plenoptic Sampling

4.2.1 IRWC Plenoptic Sampling

Our IRWC plenoptic sampling scheme can be shown in Fig. 4.12. We design a new hybrid sensing imaging system that consists of a LWIR (thermal) camera and a stereo pair of NIR cameras attached to the left and right sides of the LWIR camera. Each NIR sensor in our system samples the NIR rays and it is surrounded by a ring of LED IR lamps, which is controlled by a flash control system. With our auxiliary NIR flashes, the reflected rays from human eyes can be characterized for extremely efficient and accurate 3D eye localization. The thermal camera is used to capture reliable sources (sampling) for face identification under low light and especially under complete darkness.

For our IRWC plenoptic sampling function, we also adopt the 2PP [33] where rays are parameterized by their intersections with two parallel planes. Similar to our CWC plenoptic sampling, we assume all CoPs of the cameras lie on a common plane, the st plane with $z = 0$. For the uv plane with $z = 1$, we fix it on the image plane of cameras. We use (s, t)

to index camera position in the sampling and employ (u, v) as pixel index in each captured image. Thus, our IRWC plenoptic function is formulated as follow:

$$I = P(u_i, v_i, s_i, t_i, \lambda_i) \tag{4.16}$$

where $i = 1, 2, c$ are indices for the NIR cameras and the thermal camera.

4.2.2 Face Reconstruction

In this section, we provide a detailed face reconstruction framework based on our IRWC plenoptic sampling scheme. Fig. 4.13 shows the overall pipeline of our system consisting of 3 steps: (1) eye localization; (2) landmarks detection; (3) pose estimation and frontal face rendering. First, the two NIR cameras in the system capture two pairs of stereo images, with and without the bright-eye effect, to produce a stronger bright-eye effect which is used to localize 2D eye positions on NIR images. Then the 3D eye positions can be estimated through triangulation. We project them onto the thermal image and generates valid thermal face bounding box containing eyes. This face bounding box is further tightened by our trained thermal face detector. Second, our trained landmarks detector detects additional thermal facial landmarks in this face region. Combining projected thermal eyes' locations with the detection results, we obtain a total of 5 landmarks on the thermal image, i.e., 2 from projecting 3D eye positions and 3 directly detected from LWIR images. Finally, the head pose can be estimated based on these 5 landmarks by projecting a standard 3D head model onto the thermal image and minimizing the total projection error of the landmarks. And the frontal face can be rendered based on the Hassners method [22]. We will elaborate on each step in the following subsections.

4.2.2.1 Eye Localization

We efficiently and reliably detect the reflected infrared rays from the human eyes on NIR images based on the bright-eye effect and directly establish correspondences between ray projections of the image pair. This is different from the traditional depth estimation method, which suffers from correspondence ambiguities. With the known correspondences, we can accurately recover 3D eye locations through triangulation.

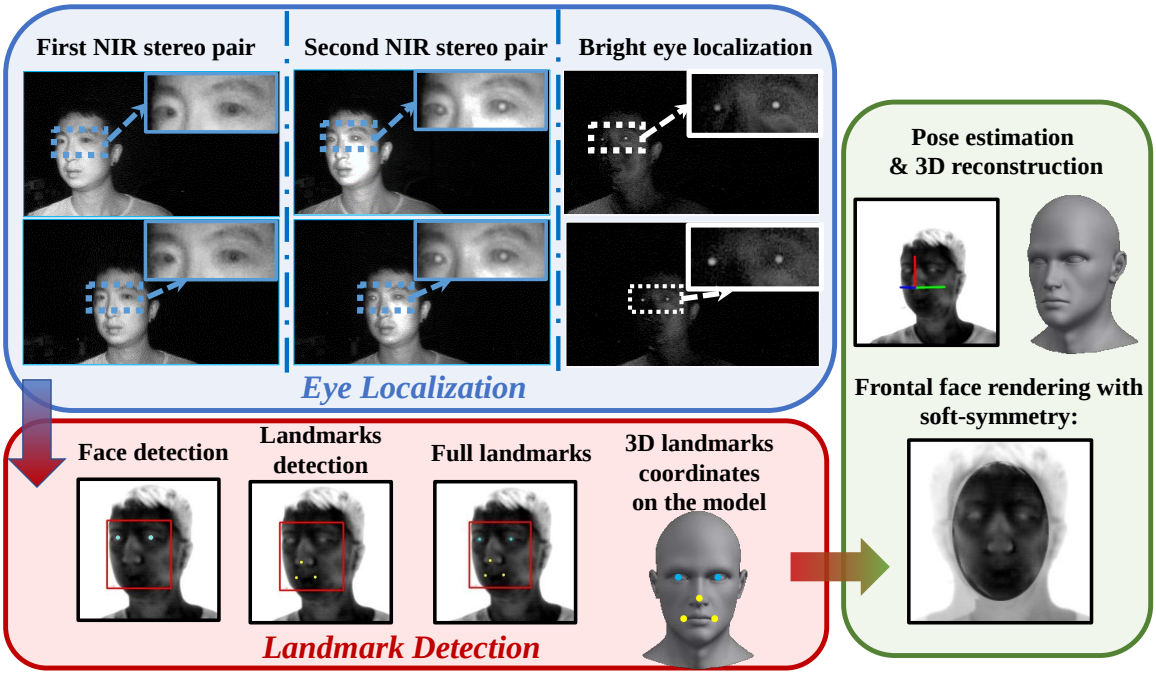


Figure 4.13: Pipeline of our face reconstruction.

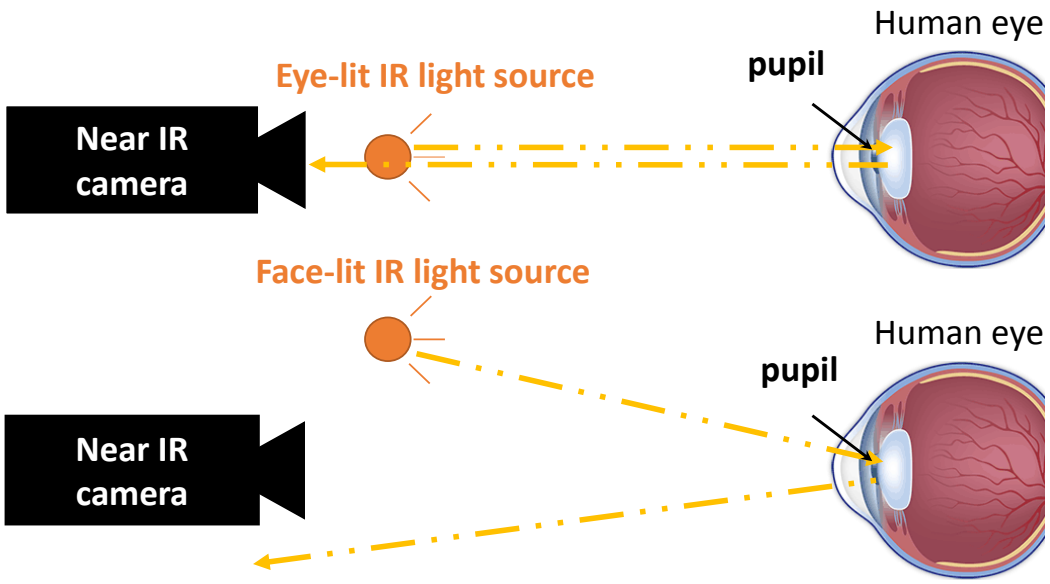


Figure 4.14: The principle of the bright-eye effect. When the NIR illuminator is on the optical axis, called the eye-lit IR light source, bright eyes are captured. When the NIR illuminator is off the optical axis, called the face-lit IR light source, no bright eyes are observed in the image.

To produce strong bright-eye effects, certain conditions need to be satisfied. The first condition is that (as discussed in [27]) the bright eyes can only be imaged if the NIR illuminator is beaming along the optical axis of the NIR camera. In such a setting, most of the light from the co-axis NIR illuminator can pass into the eye through the pupil, reflect off the fundus at the back of the eyeball, and come out through the pupil back to the image sensor. This bright-eye effect is similar to a phenomenon in photography known as the red-eye effect, except that now only the NIR part of the spectrum is captured. The bright-eye effect disappears when the NIR illuminator is positioned off the camera's optical axis, because the reflected IR light cannot enter the camera. Fig. 4.14 illustrates the underlying principle of the effects.

The second condition for strong bright-eye effects is that the light should be effectively reflective to the eyes. Behind the retina, there is ample blood in the choroid, which nourishes the back of the eyes. The blood is completely transparent for long wavelengths and abruptly starts absorbing at 600 nm [55]. Therefore, we use commodity IR light sources with wavelengths of around 800 nm, controlled by our flash control system, which employs flashes to effectively acquire the bright-eye effect with NIR cameras.

While low lighting is usually considered harmful for normal imaging situations, it actually strengthens the bright-eye effect in our NIR images. This is because pupils are fully dilated in the dark, and the IR light is minimally absorbed by the ocular pigment. An exceptional advantage of the bright-eye effect is that it is insensitive to pose variations, as shown in Fig. 4.15. Even if the subject is not facing to the optical axis of the NIR camera, the amount of IR light reflected by the retina is sufficient to produce the bright-eye effect.

To faithfully extract positions of bright eyes from each NIR camera, we instantly capture two sequential frames (within 70 ms) with and without the bright-eye effect. The first frame is the face-lit only image without bright eyes using off-axis illuminating flashes, and the second frame is the eye-lit image with bright eyes using on-axis illuminating flashes. Then, we calculate the difference map for these two frames. With this difference map, simple global thresholding will reveal the eye locations.

There are some special cases to be addressed, e.g., when the scene has other reflective

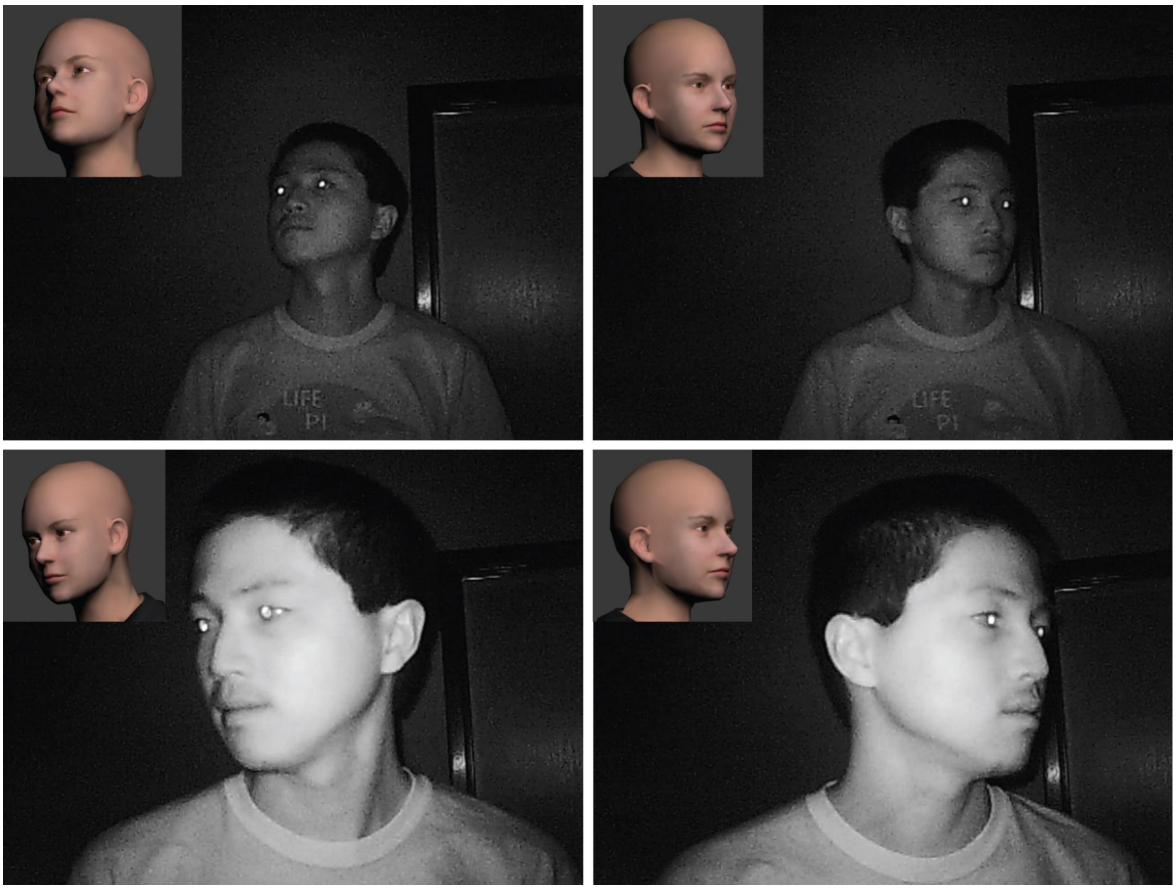


Figure 4.15: Bright-eye effects of different poses. The bright-eye effect is insensitive to pose variations under low-lighting conditions.

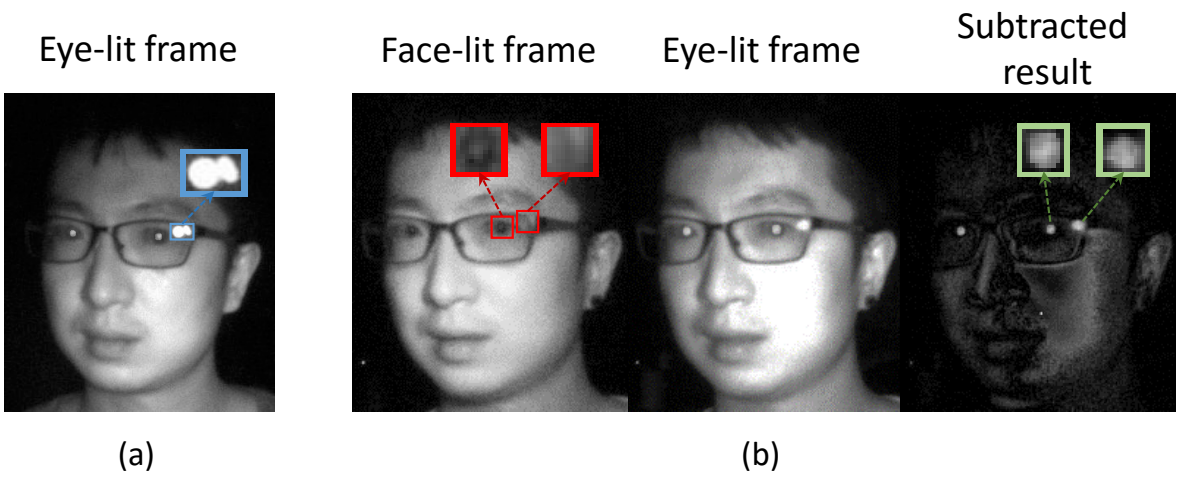


Figure 4.16: Filtering out false positives in bright-eyes' detection. (a) False positive spots have different shapes and sizes compared with the bright-eye spots, which are usually circular and small. (b) For the same shapes and sizes, eye feature patterns in the face-lit image are used to filter out false positive spots.

objects on the faces, such as glasses. To filter out false positive spots on the difference map, we use eye features to distinguish eye and non-eye objects for those potential eye spots. First, we use the shape of spots to filter out some of the false positive spots. As shown in Fig. 4.16(a), false positive spots have different shapes and sizes compared with bright-eye spots, which are usually circular and small. Second, when the spots have similar shapes, as shown in Fig. 4.16(b), we compare the features extracted from corresponding positions on the face-lit frame with standard features of true eyes to further filter out false positive spots.

After detection of eyes in the NIR stereo image pair, we recover 3D locations of eyes through triangulation and project them back to the LWIR image, forming the eye landmarks on thermal face images.

The triangulation of 3D eye locations can be conducted by solving a linear system. As illustrated in Fig. 4.17, we assume the first NIR camera coordinate system to be the world coordinate system. The projection matrix of each NIR camera is denoted as $P_i = K_i[R_i|T_i]$, where the K_i, R_i, and T_i ($i = 1, 2$, which denotes the NIR camera index) are intrinsic rotation, extrinsic rotation, and translation matrices respectively, relative to the first NIR

camera. Similarly, we define $P_c = K_c[R_c|T_c]$ as the projection matrix of the LWIR camera.

Let an unknown 3D eye coordinate be $X = (x, y, z)$ and the corresponding known homogeneous coordinates on stereo NIR images be (u_1, v_1) and (u_2, v_2), we have the following relations:

$$s_i \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = P_i \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}, i = 1, 2 \tag{4.17}$$

where s_i is an unknown scalar parameter.

The two projection matrices P_1 and P_2 for each NIR camera can be obtained through camera calibration and expressed as

$$P_i = \begin{bmatrix} p_{i,11} & p_{i,12} & p_{i,13} & p_{i,14} \\ p_{i,21} & p_{i,22} & p_{i,23} & p_{i.24} \\ p_{i,31} & p_{i,32} & p_{i,33} & p_{i,34} \end{bmatrix}, i = 1, 2 \tag{4.18}$$

By combining Eqn. 4.17 and Eqn. 4.18, and eliminating the unknowns s_1 and s_2, we derive a linear system:

$$J \begin{bmatrix} x \\ y \\ z \end{bmatrix} = b, \tag{4.19}$$

$$J = \begin{bmatrix} p_{1,11} - u_1 p_{1,31} & p_{1,12} - u_1 p_{1,32} & p_{1,13} - u_1 p_{1,33} \\ p_{1,21} - v_1 p_{1,31} & p_{1,22} - v_1 p_{1,32} & p_{1,23} - v_1 p_{1,33} \\ p_{2,11} - u_2 p_{2,31} & p_{2,12} - u_2 p_{2,32} & p_{2,13} - u_2 p_{2,33} \\ p_{2,21} - v_2 p_{2,31} & p_{2,22} - v_2 p_{2,32} & p_{2,23} - v_2 p_{2,33} \end{bmatrix}$$

$$b = \begin{bmatrix} u_1 p_{1,34} - p_{1,14} \\ v_1 p_{1,34} - p_{1,24} \\ u_2 p_{2,34} - p_{2,14} \\ v_2 p_{2,34} - p_{2,24} \end{bmatrix}$$

By solving Eqn. 4.19, we obtain unknown 3D eye coordinates as follows:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = (J^T J) \setminus (J^T b) \tag{4.20}$$

Then, we obtain the position of eyes on the thermal image (u_c, v_c) by projecting the recovered 3D eye location onto the thermal image based on calibration parameters. The pixel coordinates of eyes on the thermal image u_c, v_c can be expressed as

$$s_c \begin{bmatrix} u_c \\ v_c \\ 1 \end{bmatrix} = P_c \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{4.21}$$

where s_c is a scalar and P_c is the projection matrix of the thermal camera, which can also be obtained through calibration. We denote eye landmarks of the thermal face as $(u_{c,l}, v_{c,l})$ and $(u_{c,r}, v_{c,r})$ for the left eye (LE) and right eye (RE) respectively.

4.2.2.2 Additional Facial Landmarks Detection

In this section, we locate 3 more face landmarks to resolve the 1D ambiguity of the head pose. Besides two eye landmarks (LE and RE) on the thermal image, we detect three additional landmarks on the nose tip (N) and left and right corners of the mouth (LM and RM). We use a deep cascaded CNN to detect these landmarks on thermal face images. It's important to note that, we do not detect the additional facial landmarks on NIR images and project them onto the thermal image because the NIR images cannot provide reliable information except the bright eyes when the target is far away from the NIR cameras, as shown in Fig. 4.21.

From the identified eye landmarks identified, our system locates a potential sufficiently large face region that covers the face, where the size of the region is proportional to the target distance and the distance between the two eyes. We apply a cascade thermal face

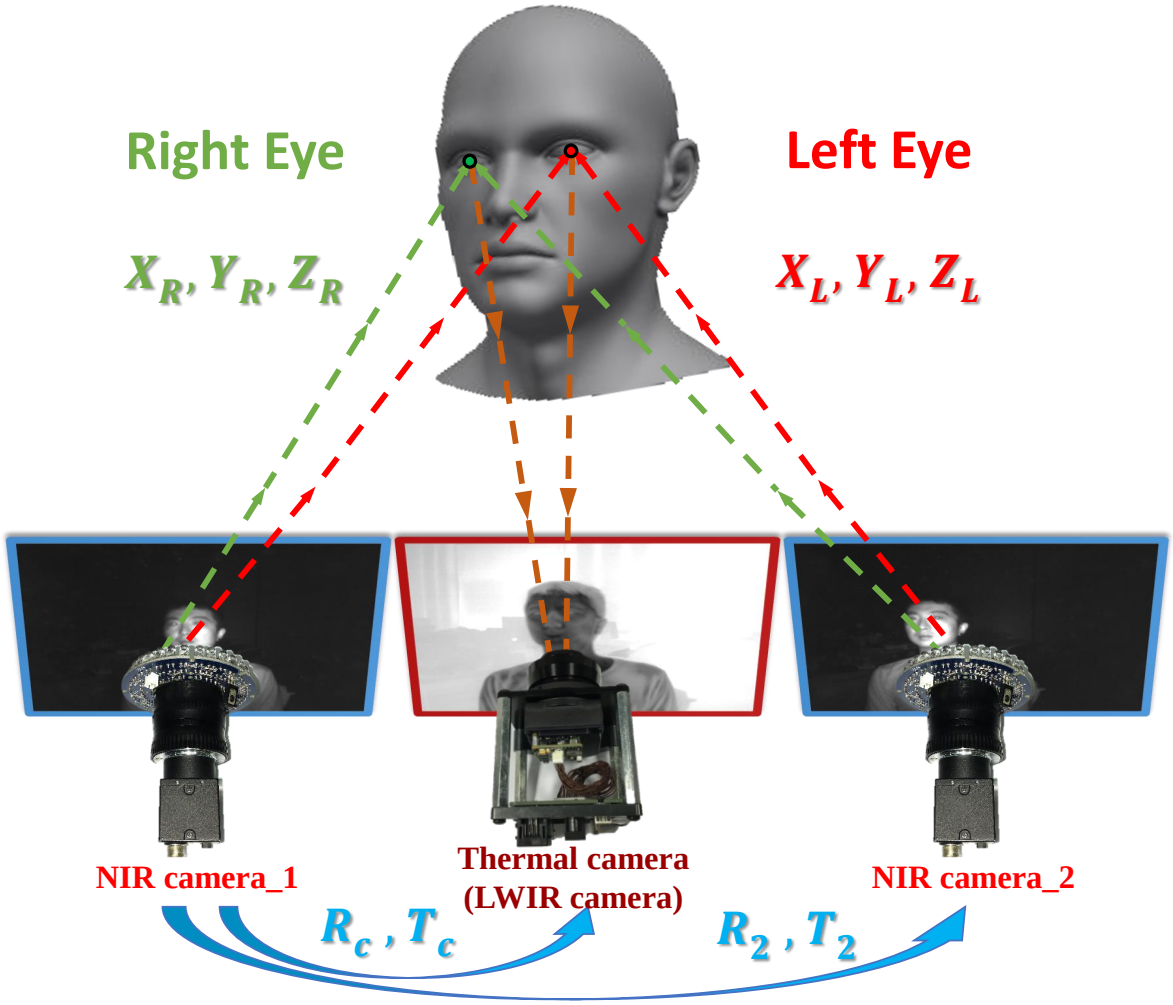


Figure 4.17: Triangulation of 3D eye locations.

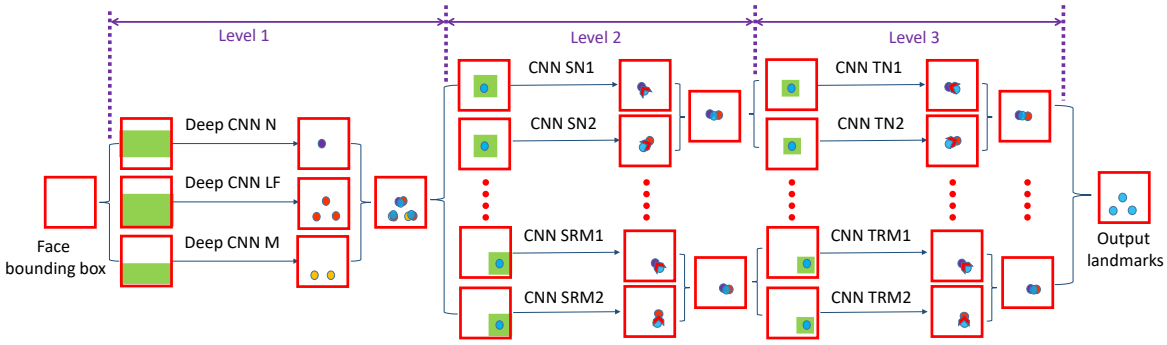


Figure 4.18: The modified cascaded CNN for thermal facial landmarks (N, LM, and RM) detection based on [51].

detector that is trained on our thermal face data set onto the potential region to tighten the face bounding box

We designed our own additional thermal face landmark detector based on Suns deep-cascaded CNN [51]. Fig.4.18 shows an overview of our modified model. Due to the weak discriminative features of the eyes in the LWIR images, we eliminate the part of the eye detection in Suns model. We adjusted three input regions in the first level of the model, which cover the nose (N), the lower face, and the mouth (M), respectively. Each deep structure in the first level is adjusted correspondingly since the sizes of the three input regions change. We eliminated the parts of the eye landmark refinement and correspondingly adjusted the size of the local search regions in the remaining levels to improve the efficiency and accuracy of our required landmarks detection, which is shown in Section 3-4.2.3.3. Since the precise face region can be predicted with the aid of the accurate eye locations, the cascade thermal facial landmarks detector is only used in the face region. Therefore, we achieve accurate detection of the three additional facial landmarks.

4.2.2.3 Pose Estimation

We perform pose estimation using 5 facial landmarks: LE, RE, N, LM, and RM. We adopt a standard 3D human head model and extract 5 corresponding standard 3D facial points. With the calibrated intrinsic and distortion parameters, we can estimate the relative pose of the head model to the thermal camera by solving a perspective-n-point (PnP)

problem. We use the method in [32] to solve this problem, where the projection error is minimized:

$$res = \sum_i dist^2(K_c[R_Q|T_Q] \begin{bmatrix} \boldsymbol{X}_i \\ 1 \end{bmatrix}, m_{c,i}) \tag{4.22}$$

where $\boldsymbol{X}_i, m_{c,i} = (u_{c,i}, v_{c,i}), i = 1, ..., 5$ are 3D points in the thermal camera coordinate system and their corresponding distortion-corrected 2D image projections on the thermal image; $dist(\boldsymbol{a}, \boldsymbol{d})$ computes the 2D distance between points $\boldsymbol{a}$ and $\boldsymbol{d}$; and K_c is the pre-calibrated thermal camera intrinsic matrix. The rotation and translation matrices of the head pose R_Q, T_Q relative to the LWIR camera are estimated by minimizing Eq. 4.22.

The accuracy of pose estimation with a single 3D shape is based on localizing facial features, as demonstrated in [22]. Failures in facial landmark detection will lead to failures in most pose estimation and correction algorithms. It is important to note: (1) our solution can provide high pose estimation and correction accuracies since we have highly accurate facial feature localizations, especially eye localizations; (2) it can still estimate a rewarding approximated pose using only two derived 3D eye positions in the case of failures in the detection of other facial features, although the estimated pose may have error (ambiguity) in one dimension (the pitch of the head pose), However, the human head does not usually have a large pitch angle, either looking up or down.

In particular, when we have two recovered 3D eye positions, we use only two pairs of 3D-to-2D points to perform pose estimation in the following. The 3D RE and LE positions are noted as $X_r = (x_r, y_r, z_r)$ and $X_l = (x_l, y_l, z_l)$ in the world coordinate system, which is the first NIR camera coordinate system. We first calculate a right vector $\vec{r} = R_c(X_l - X_r)$ and its normalized vector $\hat{r}$. We second calculate a up vector $\vec{u} = \hat{r} \times (0, 1, 0)$ and a forward vector $\vec{f} = \hat{r} \times \hat{u}$. Finally, we initialize R_Q as $[\hat{r}^T, \hat{u}^T, \hat{f}^T]$ and solve T_Q through Eq. 4.22.

4.2.2.4 Frontal Face Rendering

In this section, we perform face frontalization after pose estimation by adopting Hassner's method in [22]. First, We positions a virtual perspective thermal camera transformed by the matrix $[R_q \ T_q]^{-1}$ with respect to the original thermal camera orientation. The virtual thermal camera has plenoptic sampling $P'(u'_j, v'_j, s'_c, t'_c, \lambda_c)$ and it could capture the reference image I_r with the frontal face, shown in Fig. 4.19(c). For each pixel $m'_j = [u'_j, v'_j], j = 1, 2, ..., n$ on the frontal face of the reference image, we store the 3D point $X_j = (x_j, y_j, z_j)^T$ on the surface of the head model which is projected to m'_j.

Second, we assign a thermal intensity value for each pre-stored 3D point X_j from its projection ($m_j = [u_j, v_j]$) onto the original thermal image I_o (query image, such as the example shown in Fig. 4.19(a)) as:

$$m_j \sim K_c[R_Q|T_Q]X_j \tag{4.23}$$

$$P'(u'_j, v'_j, s'_c, t'_c, \lambda_c) = P(u_j, v_j, s_c, t_c, \lambda_c) = I_o(m_j) \tag{4.24}$$

The sampled thermal intensities $I_o(m_j), j = 1, 2, ..., n$ from bi-linear interpolation assigned to $I_r(m'_j)$ produce an initial frontalized result. Fig. 4.19(d) shows the initial frontalized view.

Third, we apply soft-symmetry processing as that in [22] to refine the initial result. Those pixels corresponding to poorly visible 3D points from the original thermal view are reassigned with intensities of their symmetric correspondence on the other side of face. We average the initial and refined output to obtain the final result. Fig. 4.19(e) shows our final frontalized face image.

4.2.3 Experiments

4.2.3.1 Camera System Prototype

Fig. 4.20 shows our prototype of the hybrid sensing system. In the center of the system is the $Tamarisk$ 640 thermal camera. We use two $FLea2$ monochrome cameras with NIR pass filters as our NIR cameras. Each NIR camera is surrounded by a $Logisaf$

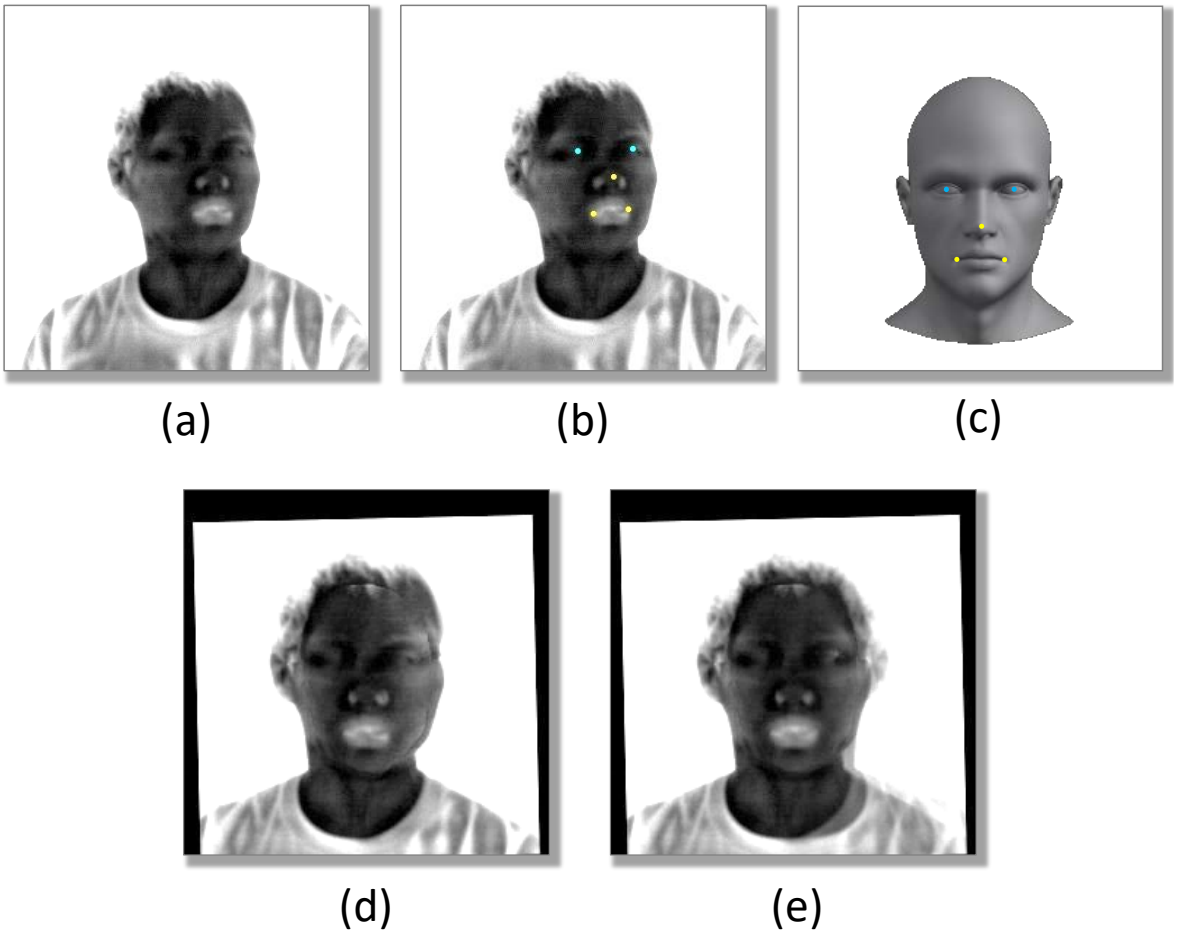


Figure 4.19: Frontal face rendering overview. (a) Query image. (b) Fused landmarks image. (c) Reference view with detected landmarks rendered from the 3D head model, each of its pixels on the face has corresponding 3D point coordinate located on surface of the 3D model. (d) Back-project query intensities to the reference coordinate system. (e) Frontalized result with soft-symmetry.

L002-48-94 IR board (940 nm), which is a component of the $Logisaf$ CCTV camera, to simulate the on-axis IR light source, i.e., eye-lit IR lights. We also place two extra $Logisaf$ L002-60-94 IR boards (940 nm) above the two NIR cameras as the off-axis light source, i.e., face-lit IR lights.

The first challenge in our hybrid system is the synchronization of IR illuminators, the pair of NIR cameras and the LWIR camera for the image acquisition. We synchronize the pair of NIR cameras through the camera operation APIs provided by Point Grey. However, it is more challenging to synchronize the on-axis and off-axis IR light sources with the camera. For this purpose, we design an IR flash control system (shown on the left of Fig. 4.20) comprising one Phidget Interfacekit board and two relay boards. This system not only synchronizes our IR light sources with the camera system but also turns these sources into flashes, which makes our system more covert compared with continued NIR illumination. The control system is programmed to first turn on off-axis IR lights; simultaneously trigger NIR cameras to capture the first frame (without bright eyes) and turn off these lights; and then turn on on-axis IR lights, trigger the cameras for the second frame (with bright eyes), and turn them off. The working range of this prototype is shown in Fig. 4.21.

4.2.3.2 Calibration

The another challenge is the cross-modality cameras calibration, since the printed chessboard or other pattern is not visible due to its uniform temperature in the LWIR image. To efficiently calibrate the LWIR camera, we design a white pattern mold (shown in Fig. 4.22(a)) with circular holes on the board. We keep the mold at a normal temperature and use a black exothermic board behind the mold to produce a higher temperature than that of the board. This leads to a clear contrast between the circular holes on the mold and the mold in the LWIR image (shown in Fig. 4.22(c)). After extracting centers of these circles in the pattern, we use a calibration code [73] for LWIR camera calibration. Due to the color contrast between the black board and the white mold, it is easy to detect centers of the circles on captured images from Flea2 cameras, the example of which is shown in Fig. 4.22(b). Ultimately, we solve the calibration problem with this specially designed mold.

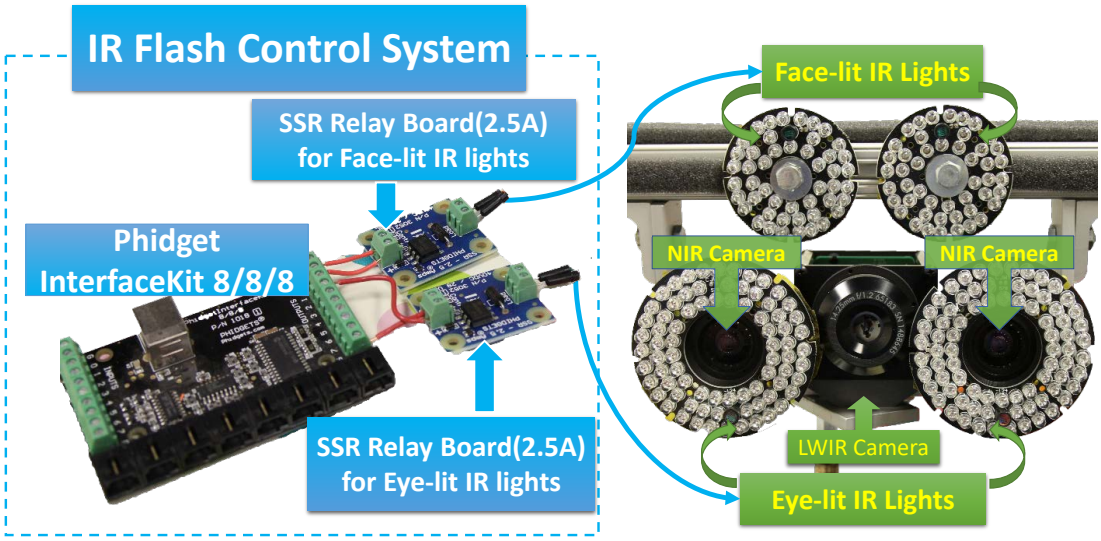


Figure 4.20: Our camera system prototype.

Table 4.1 shows intrinsic calibration results for the pair of NIR cameras and the thermal camera, K_1, K_2, and K_c. Parameters (f_x, f_y) represent camera focal lengths, (c_x, c_y) represent optical centers expressed in pixels coordinates, $k_1, k_2, ..., k_6$ represent radial distortion coefficients (k_4, k_5 are fixed to 1 in calibration), and p_1 and p_2 represent tangential distortion coefficients. RMS in the last row of Table 4.1 illustrates root-mean-squared distances in pixels between detected image points and projected ones. Table 4.2 shows extrinsic calibration results, R_2, T_2, R_c, and T_c, and the re-projection error for different stereo pairs of cameras. This error is also calculated by RMS for all points in all the available views from each stereo pair. The small calibration errors, both in intrinsic and extrinsic calibration result tables, indicate that our calibration method with the designed tool is accurate and reliable.

4.2.3.3 Face Reconstruction Results

Results of Landmark Detection. In this section, we compare the performance of landmarks detection by the method in [51] and by our method. We acquire a small data set using our hybrid sensing system to demonstrate our method. We randomly select 65% of the data

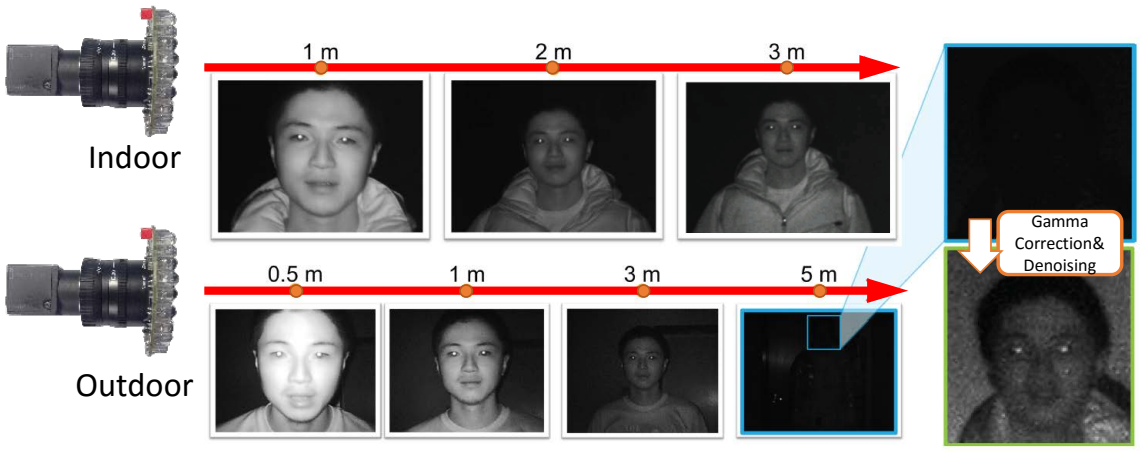


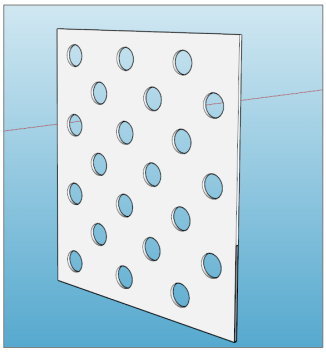
Figure 4.21: The working range of our prototype.

Parameters	NIR Camera_1	NIR Camera_2	Thermal Camera
c_x	311.2695	302.0133	322.0687
c_y	258.8132	276.4582	246.3135
f_x	1046.5966	1055.7334	841.6957
f_y	1047.2157	1056.2581	840.2238
k_1	-0.1241	-0.2270	-0.3813
k_2	-3.4701	1.4192	-0.5758
p_1	0.0018	0.0024	-0.0026
p_2	-0.0012	0.0004	-0.0009
k_3	-998.1297	572.7115	-70.6259
k_6	-1046.1720	595.1688	-79.4948
RMS	0.2017	0.1810	0.3015

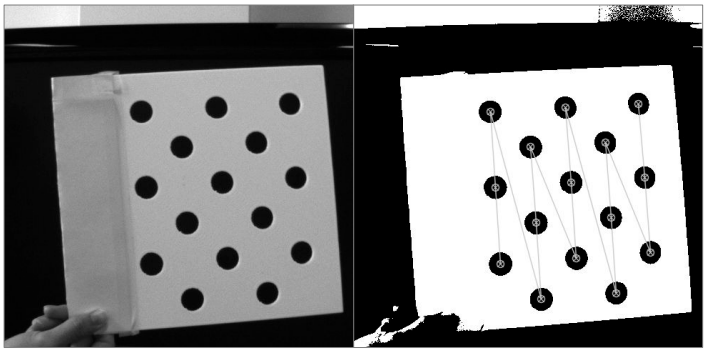
Table 4.1: Intrinsic calibration result.

set for training, 10% for validation, and the remaining images for testing. The training input to the learning-based method in [51] requires labeled eye landmarks. However, it is very difficult to manually label eye landmarks in thermal images due to lack of information around the eyes. This can be seen in Fig. 4.25(a), where the eye region of the thermal image is compared with the eye region of the NIR image.

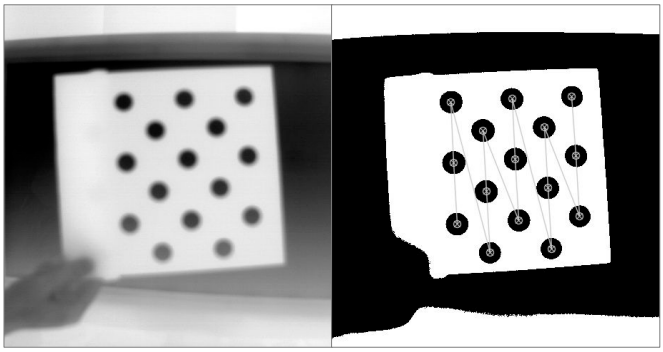
We use projected eye locations in thermal images as eye landmarks together with three labeled facial landmarks (N, LM, and RM) to train Sun's structure in [51]. This is



(a) Calibration mold



(b) NIR image and its thresholded image



(c) LWIR image and its thresholded image

Figure 4.22: Cross-modality camera calibration.

Pairs	Rotation Vector	Translation Vector	Re-projection Error
NIR_Cam_1 *to* NIR_Cam_2	(-0.01237; 0.00816; -0.01713)	(113.34144; 0.42439; 2.83881)	0.19021
NIR_Cam_1 *to* $LWIR\_Cam$	(-0.00629; -0.00042; -0.00852)	(55.06529; -0.79071; -5.88902)	0.28276

Table 4.2: Extrinsic calibration result.

the best scenario of training Sun's structure for detection of all the five thermal facial landmarks. For each testing data, these five "labeled" facial landmarks are used as ground truth to compare with detected landmark results.

We use the same performance measurements in [51] to calculate the average detection error and the failure rate for each facial point. These measurements indicate the accuracy and reliability of the algorithm. The detection error is measured as:

$$err = \sqrt{(u_t - \tilde{u}_t)^2 + (v_t - \tilde{v}_t)^2}/l \tag{4.25}$$

where (u_t, v_t) and $(\tilde{u}_t, \tilde{v}_t)$ are the ground-truth position and the detected position, and l is the

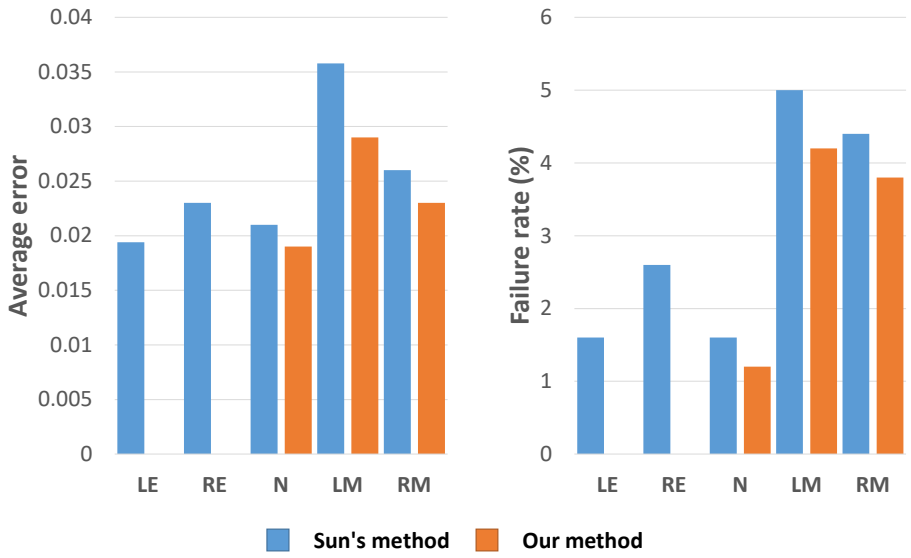


Figure 4.23: Average detection errors and failure rates of the Sun's structure [51] and ours on the testing data.

width of the face bounding box returned by the thermal face detector. A failure is counted if an error is larger than 5%.

Fig. 4.23 summarizes the comparison results of landmarks detection by the learning-based method in [51] and by our method on the testing data. Our method achieves higher accuracy for detection of both thermal eyes and other thermal facial landmarks (N, LM, and RM). More than 9.5%, 18.9% and 11.5% relative accuracy improvements on average errors for N, LM and RM respectively. Besides, our thermal eyes detection has no failure, training and testing error. For the learning-based method, lack of discriminative information around the eyes (low contrast) in thermal images causes inaccurate eye detection and less accurate detections of the other three facial landmarks.

Compared with the learning-based method [51] that only uses thermal images, our method has two advantages for achieving more accurate facial landmark detection. First, it locates eye landmarks in thermal images more accurately from 3D eye positions obtained by two NIR images in our solution. Second, it predicts the face bounding box more precisely based on the information of thermal eye positions and the actual distance between eyes in 3D, improving detection of the other three facial landmarks (N, LM, and RM) in the thermal

images.

Fig. 4.24 shows an example of comparison between our face landmark detection and sun's method [51]. Fig. 4.24(a) shows the result of our fused eye landmarks and three more facial landmark detection. Fig. 4.24(b) shows all the five facial landmark detection results using [51]. Fig. 4.24(c) shows the eye landmark location comparison.

Detection Results with Eyeglasses. We also test our eye landmark detection method on images with eyeglasses; examples are shown in the first row of Fig. 4.28. This is another advantage of our method over the learning-based method for eye detection on a thermal image. Our system uses extra NIR images, which are more informative in cases such as targets wearing eyeglasses. From Fig. 4.25(b), we can see that glasses completely block eyes on thermal images. In contrast, they have little effect on NIR images on which our eye localization is performed.

Pose Estimation and Frontalization Results. Fig. 4.26, 4.27 and 4.28 present examples of several scenarios. Fig. 4.26 shows the results of face images with horizontal rotation angles between 0° and 60° and vertical rotation angles between -45° and 45°. The first row shows landmark detection results. The second row shows pose estimation results. The third row shows pose correction results. Fig. 4.27 shows the results of face images with horizontal rotation angles between -60° and 0° and vertical rotation angles between -45° and 45°. Fig. 4.28 shows the results of face images with glasses. These results demonstrate that our method is effective and robust in head pose estimation.

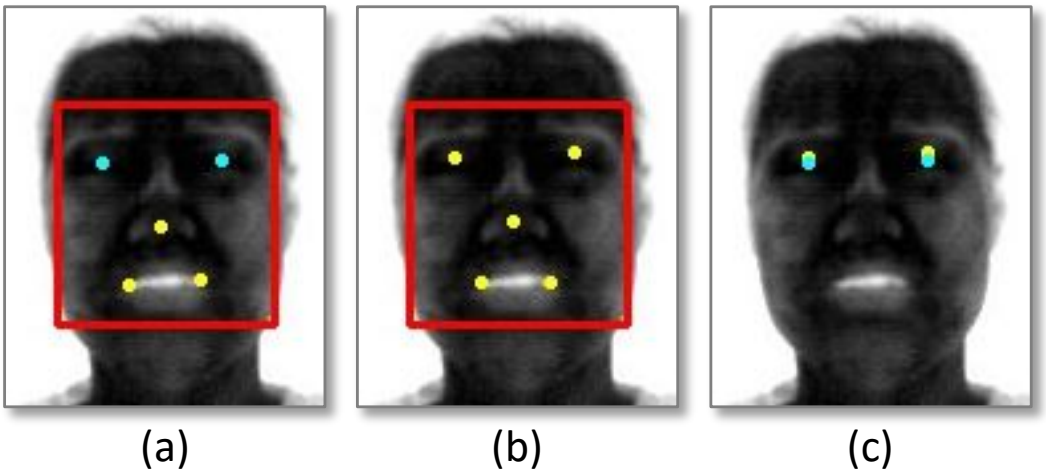


Figure 4.24: Comparison between CNN eye detection [51] (yellow points) and our method (cyan points). (a) Our fused result. (b) Full landmarks detect result from [51]. (c) The comparison.

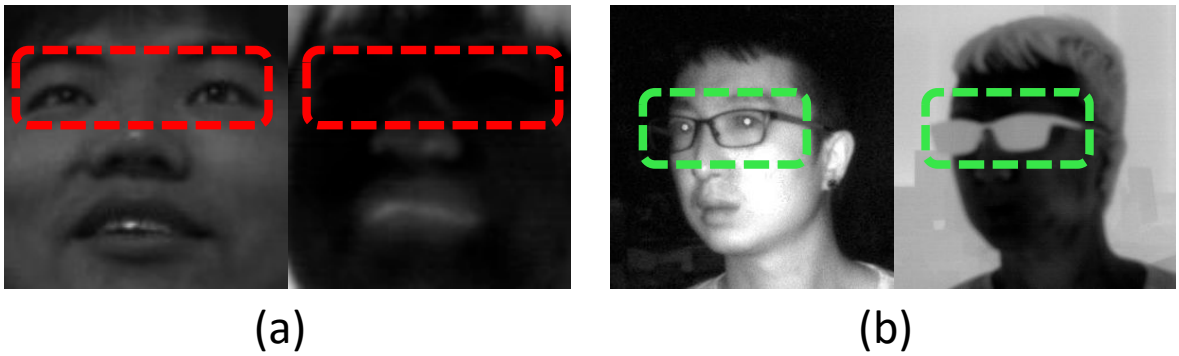


Figure 4.25: Comparison between the NIR image and thermal image. (a) Eyes region comparison. (b) Comparison with the effect of eye-glasses.

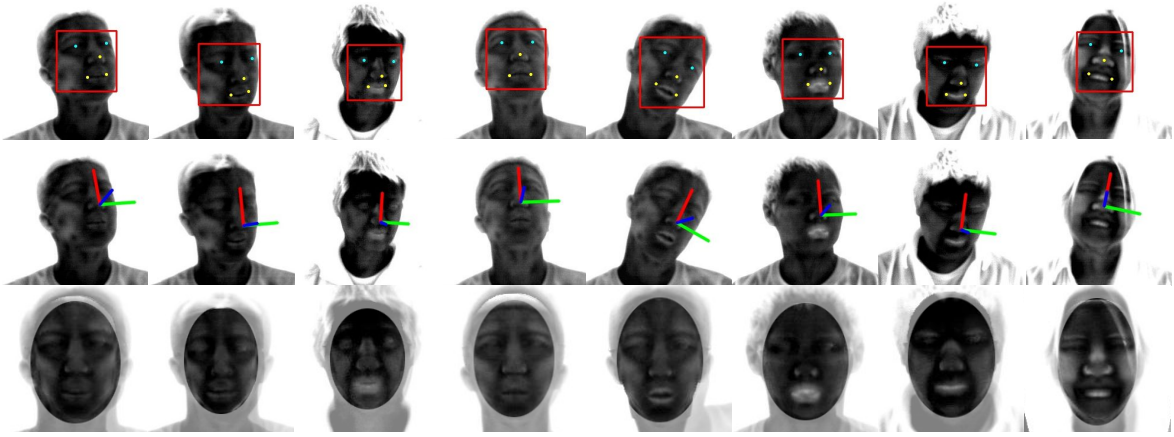


Figure 4.26: Experimental results with horizontal rotation angles between $0°$ and $60°$ and vertical rotation angles between $-45°$ and $45°$. The first row shows landmark detection results, the second row shows pose estimation results, and the third row shows frontal face rendering results.

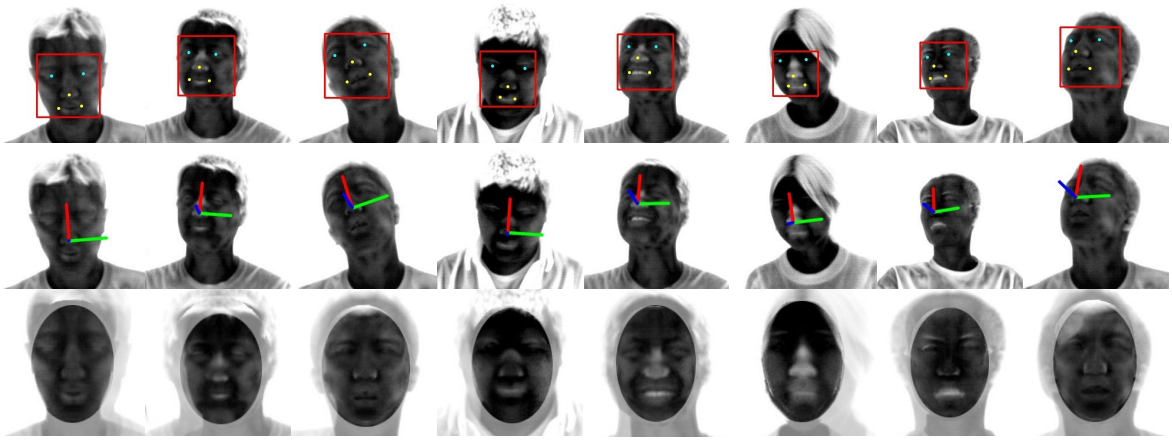


Figure 4.27: Experimental results with horizontal rotation angles between $-60°$ and $0°$ and vertical rotation angles between $-45°$ and $45°$. The first row shows landmark detection results, the second row shows pose estimation results, and the third row shows frontal face rendering results.

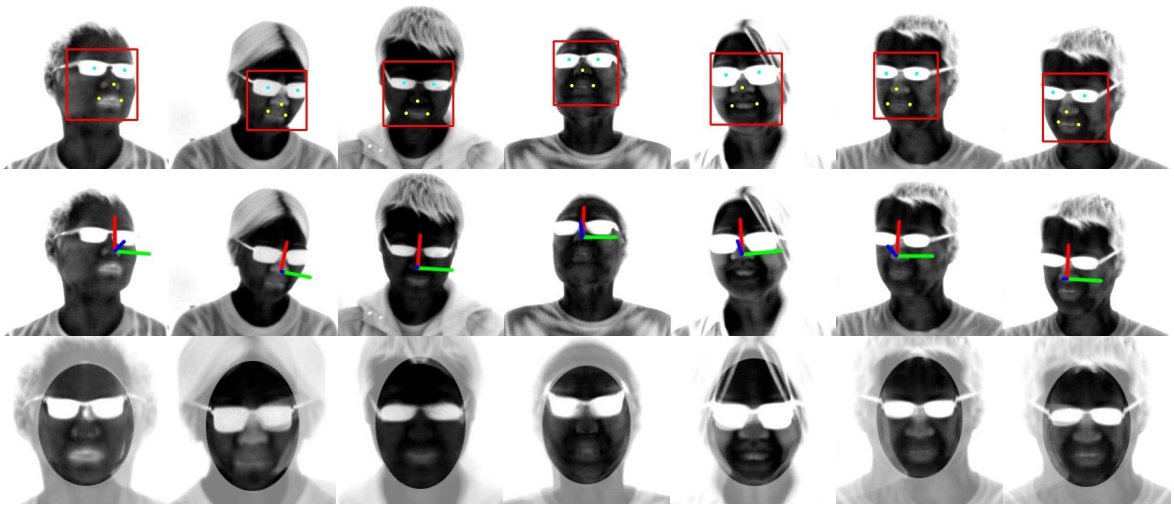


Figure 4.28: Experimental results with eye glasses appearances.

Chapter 5

SHAPE RECOVERY FROM POLARIMETRIC PLENOPTIC SAMPLING

Shape from polarization(SfP) employs the Fresnel phenomenon that unpolarized light is partially polarized after being reflected by surfaces. The polarization images can be acquired by rotating a polarizer in front of a camera, the captured varying radiance can help us infer surface normal. Such SfP techniques have three advantages. First, unlike the traditional photometric stereo, SfP does not require controlled lighting conditions, and is applicable to outdoor scenes. Second, it can handle a large variety of surface reflectances, ranging from dielectrics to metals and the translucent. Last and perhaps the most important, it bypasses feature correspondence matching and can robustly handle featureless and transparent objects.

We present a polarimetric plenoptic sampling scheme and explore a novel framework for surface reconstruction based on our sampling. Theoretically, we derive a comprehensive theory correlating the polarization radiance function with both surface normal and depth. Based on this derivation, we extend the shape-from-motion theory by viewing our plenoptic spatial sampling as a moving camera and get a new formulation under our polarimetric sampling for shape reconstruction. In particular, we prove that the reconstruction framework effectively resolves the azimuth-zenith ambiguity by forming an over-constrained (non-linear) system.

5.1 Basics on Polarization and Reflection

We first quickly review the basics on polarization. When unpolarized light is reflected from the surface, it becomes partially polarized. There are three parameters [60] completely determine the state of this partial linear polarization: intensity, phase angle, and degree of polarization (DoP). The measured intensity of the transmission of a linearly polarized light

wave passing through a polarizer with the polarizing angle α is given by the transmitted radiance sinusoid (TRS) as:

$$I_p = \frac{I_{max} - I_{min}}{2} \cos(2 * (\alpha - \varphi)) + \frac{I_{max} + I_{min}}{2} \tag{5.1}$$

where I_{max} and I_{min} are maximal and minimal magnitudes passing through the polarizer, and φ is the phase angle. Existing SfP method obtains the phase angle and DoP ρ by decomposing the TRS function. From the TRS equation, we find the phase angle has $\pi/2$ ambiguity. The DoP describes how much the light has been polarized. It is equal to 1 for perfectly polarized light and 0 for unpolarized light. ρ can be computed as:

$$\rho = \frac{I_{max} - I_{min}}{I_{max} + I_{min}} \tag{5.2}$$

Most surfaces exhibit three types of polarized reflection [7, 16]: polarized specular reflection, polarized diffuse reflection and unpolarized diffuse reflection. When illuminated using unpolarized lights, the former two can potentially recover surface normals. By applying Fresnel equations, the DoPs in terms of Fresnel coefficients for diffuse polarization and specular polarization are:

$$\rho_d = \frac{(n - \frac{1}{n})^2 \sin^2(\theta)}{2 + 2n^2 - (n + \frac{1}{n})^2 \sin^2(\theta) + 4\cos(\theta)\sqrt{n^2 - \sin^2(\theta)}} \tag{5.3}$$

$$\rho_s = \frac{2\ \sin^2(\theta)\cos(\theta)\sqrt{n^2 - \sin^2(\theta)}}{n^2 - (1 + n^2)\sin^2(\theta) + 2\sin^4(\theta)} \tag{5.4}$$

where θ is the incident angle, n is the relative refractive index ranging from 1.4 to 1.6 for dielectrics. In our technique, we assume $n = 1.5$ and later show that smaller variations on n will not affect the final estimation since it has a slight effect on the DoP. For diffuse polarization, the DoP has a one-to-one mapping of the incident angle. In contrast, for specular polarization, the θ ambiguity is shown in Fig. 5.1.

Previous studies [7] have shown that diffuse and specular polarized reflections have a $\pi/2$ difference in the phase angle φ. Combining Eqn. 5.1 and Eqn. 5.2, we can rewrite the measured intensity function for diffuse and specular polarizations as:

$$I_p^{(d)} = I_{dp} * (\frac{\rho_d}{2} \cos(2 * (\alpha - \varphi)) + \frac{1}{2}) \tag{5.5}$$

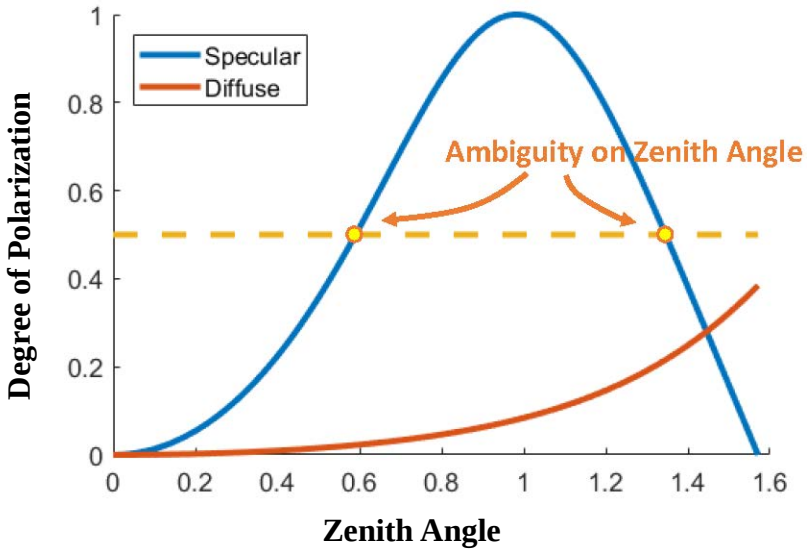


Figure 5.1: Zenith Angle vs. Degree of Polarization for specular and diffuse polarization, note that there is an ambiguity on zenith angle for specular polarization.

$$I_p^{(s)} = I_{sp} * (\frac{\rho_s}{2} \cos(2 * (\alpha - \varphi + \frac{\pi}{2})) + \frac{1}{2}) \tag{5.6}$$

where $I_{sp} = (I_{max}^{sp} + I_{min}^{sp})$ and $I_{dp} = (I_{max}^{dp} + I_{min}^{dp})$ are unpolarized intensities for diffuse and specular regions respectively.

5.2 Polarimetric Plenoptic Sampling

Our polarimetric plenoptic sampling can be shown in Fig. 5.2. First, We adopt the two-plane parameterization (2PP) [33] where rays are parameterized by their intersections with two parallel planes. We assume all CoPs of the sampling cameras lie on a common plane, the st plane with $z = 0$. We fix the uv plane with $z = 1$ on the image plane of cameras. Therefore, we use $P = (s, t)$ to index camera position in the sampling and employ (u, v) as pixel index in each captured image. For each camera position (s, t), we attach a polarizer in front of the lenses at a specific polarizing angle α_i. We use $D_i = [\cos\alpha_i, \sin\alpha_i, 0]$ as a directional vector parallel to the uv/st plane to specify the polarizing angle. Our polarimetric

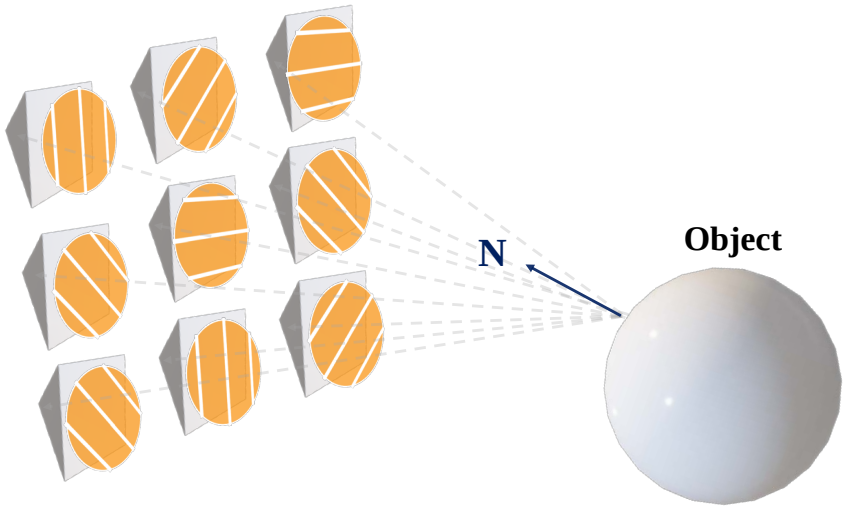


Figure 5.2: Polarimetric plenoptic sampling.

plenoptic function is formulated as follow:

$$I_p = P(u, v, s, t, \alpha) \tag{5.7}$$

Then we derive a comprehensive theory that correlates the polarization radiance function with both surface normal and depth under our polarimetric plenoptic sampling.

5.2.1 Polarization Radiance Function

For our polarimetric plenoptic sampling, we assume that the center view is at $P_c = [0, 0, 0]^T$ and its coordinate system aligns with the world coordinate system. Every other viewpoint i lies at $P_i = P_c + \boldsymbol{\tau}_i$ where $\boldsymbol{\tau}_i = (\tau_x^i, \tau_y^i, 0)^T$. Consider a 3D point $X = [x, y, z]^T$ with its surface normal as $N = [n_x, n_y, n_z]^T$. The viewing direction towards the point at camera P_i is $V_i = P_i - X$.

The incident angle θ between the light direction and surface normal is equal to the angle between the surface normal N and viewing direction V_i, and its trigonometric functions can also be determined by these two terms N and V_i as $\cos(\theta) = (N \cdot V_i)/\|V_i\|$ and $\sin(\theta) =$

$\|N \times V_i\|/\|V_i\|$. By applying $\cos(2a) = 2\cos^2(a) - 1$, we can model the terms $\cos(2(\alpha - \varphi))$ and $cos(2 * (\alpha - \varphi + \frac{\pi}{2}))$ in polarized intensity I_p with N, V_i and D_i as:

$$\kappa_d = \cos(2 * (\alpha - \varphi)) = 2(D \cdot \frac{((N \times V_i) \times V_i)_{xy}}{\|((N \times V) \times V)_{xy}\|})^2 - 1 \tag{5.8}$$

$$\kappa_s = cos(2 * (\alpha - \varphi + \frac{\pi}{2})) = 2(D \cdot \frac{(N \times V)_{xy}}{\|(N \times V)_{xy}\|})^2 - 1 \tag{5.9}$$

Where $(C)_{xy} = [C_x, C_y, 0]^T$. Combining Eqs. 5.3 5.4, 5.5 and 5.6, we can derive two new transmitted radiance sinusoid functions in terms of N, V, n and $I_{sp/sd}$ for specular and diffuse polarizations as:

$$I_p^{(d)}(N, V, D) = I_{dp} * (\frac{\kappa_d (n - \frac{1}{n})^2 \|N \times V\|^2}{(2 + 2n^2)\|V\|^2 - (n + \frac{1}{n})^2 \|N \times V\|^2 + 4(N \cdot V)\sqrt{n^2\|V\|^2 - \|N \times V\|^2}} + 1) \tag{5.10}$$

$$I_p^{(s)}(N, V, D) = I_{sp} * (\frac{2\kappa_s (N \cdot V)\|N \times V\|^2 \sqrt{\|V\|^2 n^2 - \|N \times V\|^2}}{n^2\|V\|^4 - (1 + n^2)\|V\|^2\|N \times V\|^2 + 2\|N \times V\|^4} + 1) \tag{5.11}$$

Our derivation shows that polarization image radiance functions are only related to the surface normal, depth and the intensity. Thus, we have:

$$P(u, v, s, t, \alpha) = I_p(N, V(u, v, s, t), D(\alpha)) \tag{5.12}$$

5.2.2 Differential Analysis

To analyze how the polarization image changes according to the viewpoint, we make a differential analysis. From the analysis, we derive a relation that relates camera motion to surface normal and depth.

For a 3D point X, assume that it is projected to pixel $\mathbf{u} = (u, v)^T$ for the center perspective camera:

$$u = \frac{x f_x}{z} + c_x, \qquad v = \frac{y f_y}{z} + c_y \tag{5.13}$$

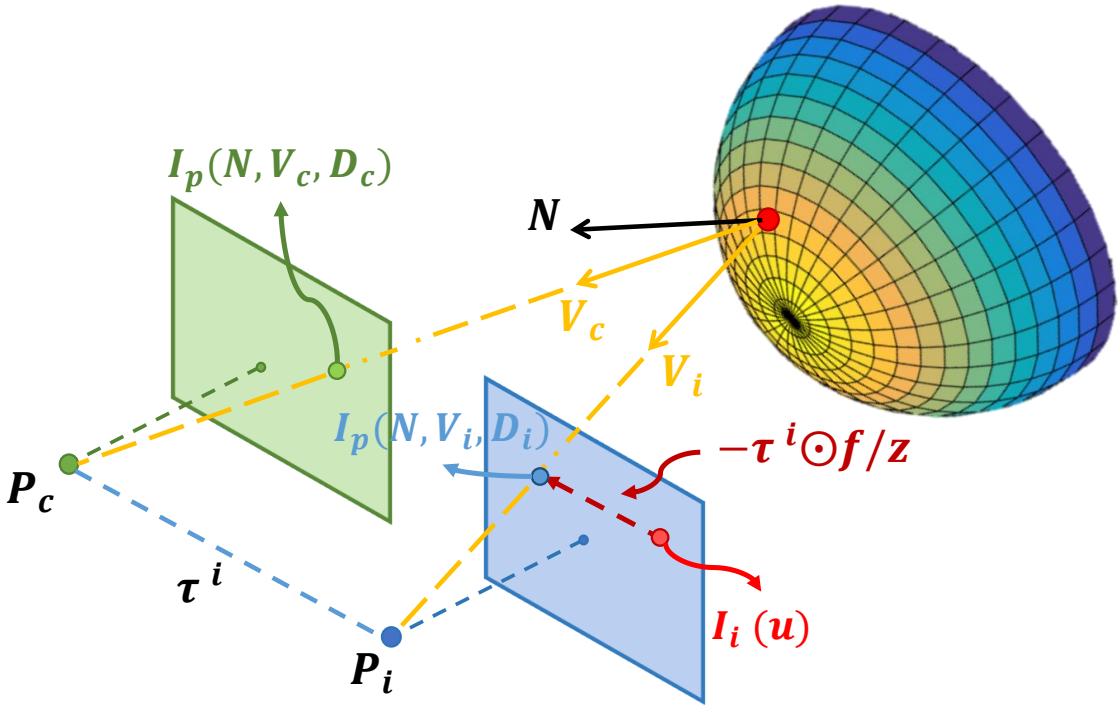


Figure 5.3: Optical flow for the surface.

where $\mathbf{f} = (f_x, f_y)^T$ are local lengths of the camera, and $\mathbf{c} = (c_x, c_y)^T$ are principle points. Let V_c and D_c be viewing directions from X to the center camera and polarizing directional vector at center, as shown in Fig. 5.3.

For the viewpoint P_i, the projected pixel position of X is $\mathbf{u} + \delta\mathbf{u}_i$ with the intensity $I_p(N, V_i, D_i)$. The camera movement can be viewed as moving the object by $\delta X_i = -\boldsymbol{\tau}_i$. We have $\delta\mathbf{u}_i = \delta X_i \odot \mathbf{f}/z = -\boldsymbol{\tau}_i \odot \mathbf{f}/z$. Therefore, we derive the relation analogous to the optical flow:

$$I_p(N, V_i, D_i) \cong I_i(\mathbf{u}) + (\bigtriangledown_{\mathbf{u}} I_i)^T \delta\mathbf{u}_i \tag{5.14}$$

where $(\bigtriangledown_{\mathbf{u}} I_i)^T = (I_u^i, I_v^i)^T$ is spatial derivatives calculated from image I_i. By expanding derivatives in the Eqn. 5.14, we get:

$$I_p(N, V_i, D_i) + I_u^i \frac{\tau_x^i f_x}{z} + I_v^i \frac{\tau_y^i f_y}{z} - I_i(\mathbf{u}) = 0 \tag{5.15}$$

where V_i is determined by depth z. We observe that the equation above only contains N, z, and I_{sp} (or I_{dp}).

5.3 Shape Recovery

Next, we show how to use above analysis to recover surface normal and depth. Recall each view camera obeys the relation of Eqn. 5.15. We therefore can stack all these equations to form a system of nonlinear equations. Solving this system simultaneously corresponds to recovering the surface normal and depth. Further, if we have $\tau_i = 0$ for all equations, equivalent to all polarized images being captured at the central view, solving surface normal from the Eqn. 5.15 corresponds to the traditional SfP.

Recall that we aim to solve a highly non-linear system, it is therefore essential to obtain good initialization for both surface normal and depth. We apply the traditional light-field depth estimation method [15] on the views with same polarization angles to first obtain a coarse depth map, and then generate a coarse normal map based on it.

Surface normal initialization also enhances disambiguation for the azimuth angle. For an unknown unpolarized intensity $I_{sp/dp}$, we use the initialized depth map to warp two polarized images with a 90-degree angle difference between the polarizer and the center view. Next, we initialize its unknown unpolarized intensity by adding up the two polarized images: $I_{sp/dp} = I_p(D(\alpha)) + I_p(D(\alpha + 90))$.

We can fuse normal and depth via the following optimization. Our target function consists of a data term E_d and a smoothness term E_s as:

$$\underset{N,z}{\operatorname{argmin}} \sum_S E_d^2 + \eta \sum_S E_s \tag{5.16}$$

where S is the 4-neighbor patch, η the weight, E_d calculated by the left hand side of Eqn. 5.15, and $E_s = \|N - \bar{N}\|^2 + \|z - \bar{z}\|^2$ where $\bar{N}$ and $\bar{z}$ are the averages of surface normal and depth of neighbors.

Once we get the initialization, we apply the Levenberg-Marquardt method to optimize the surface normal and depth. In order to make above optimization converging to the global solution, after the first optimization, we solve the optimization again initializing with the π differences in the recovered phase angles, which can correct the recovered shape. Note that, we do not know each measured radiance whether the polarized diffuse reflection dominates

or the polarized specular reflection dominates, therefore, we set the data term with diffuse and specular polarization radiance functions respectively and find the optimal results as the final results.

We observe that viewing direction variations in the sampling further help resolve the azimuth-zenith ambiguity. This is because the azimuth-zenith ambiguity can produce at most four surface normal candidates at each viewpoint whereas the ground truth can be easily derived from multiple viewpoints as the common candidate. Theoretically, 4 cameras are enough to recover 3D shape, but, in practice we use 5×5 camera array to ensure a robust performance.

5.4 Experiments

In this section, We first talk about our acquisition system construction. Next, we validated our reconstruction approach on both synthetic and real data. All experiments are conducted on a desktop with an Intel i7 7820 CPU (2.9GHz Quad-core) and 32G memory. Our algorithm is implemented in the Matlab.

5.4.1 Camera System Construction

We construct a polarimetric camera array to evaluate our approach on real data. To construct the camera array, we mount a polarization camera (Phoenix PHX050S-P/Q) with 8.5mm lens on a translation stage to uniformly translate the camera position on a 2D plane, as shown in Fig. 5.4. We use the polarization camera instead of rotating a polarizer for polarimertic sampling. LUCIDs Phoenix camera has the IMX250MZR Polarsens sensor which incorporates a layer of polarizers, and the polarizer array layer has four different angled polarizers (90^o,45^o,135^o, and 0^o). For each sampling camera position, we extract the image with respect to one polarization angle from the raw image through the demosaicing process. With this setup, we can acquire our polarimetric plenoptic sampling. The camera resolution is 2448×2048. We put observed objects into a photo studio kit box. The distance between the objects and the acquisition system is about 70 cm.

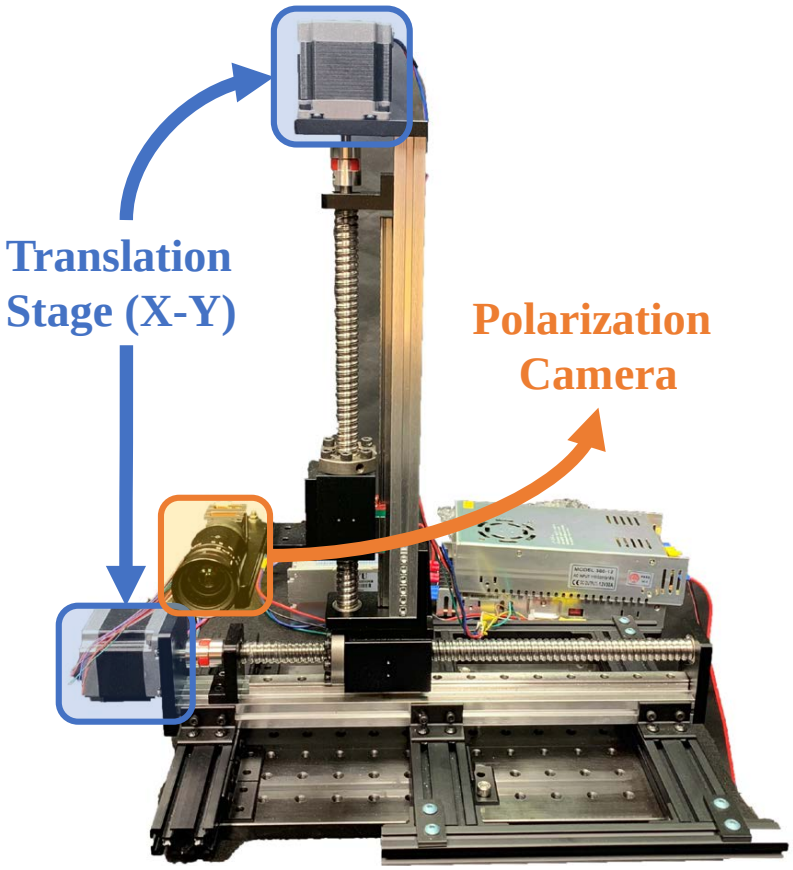


Figure 5.4: Polarimetric plenoptic sampling acquisition system setup.

5.4.2 Synthetic Result

We first generate a polarization renderer to render the polarization images. We evaluate our method on a simple synthetic sphere scene where the radius of the sphere is 5 and the distance between the camera array and object is 30. The resolution of each view is 400 by 400, and the baseline between neighboring cameras is 0.1. We capture a 5x5 plenoptic sampling. we render two sphere scenes with polarized diffuse reflection and polarized specular reflection respectively, The first column in the Fig. 5.5 shows the sample images from our polarimetric plenoptic sampling and the recovered surfaces are shown in the second column, the last column indicates the surface normal error map, we can see that all the degree errors are kept within 5 degrees. Moreover, for both diffuse and specular polarization scenes, the estimated surface normal results do not contain the large errors caused by the azimuth-zenith ambiguity.

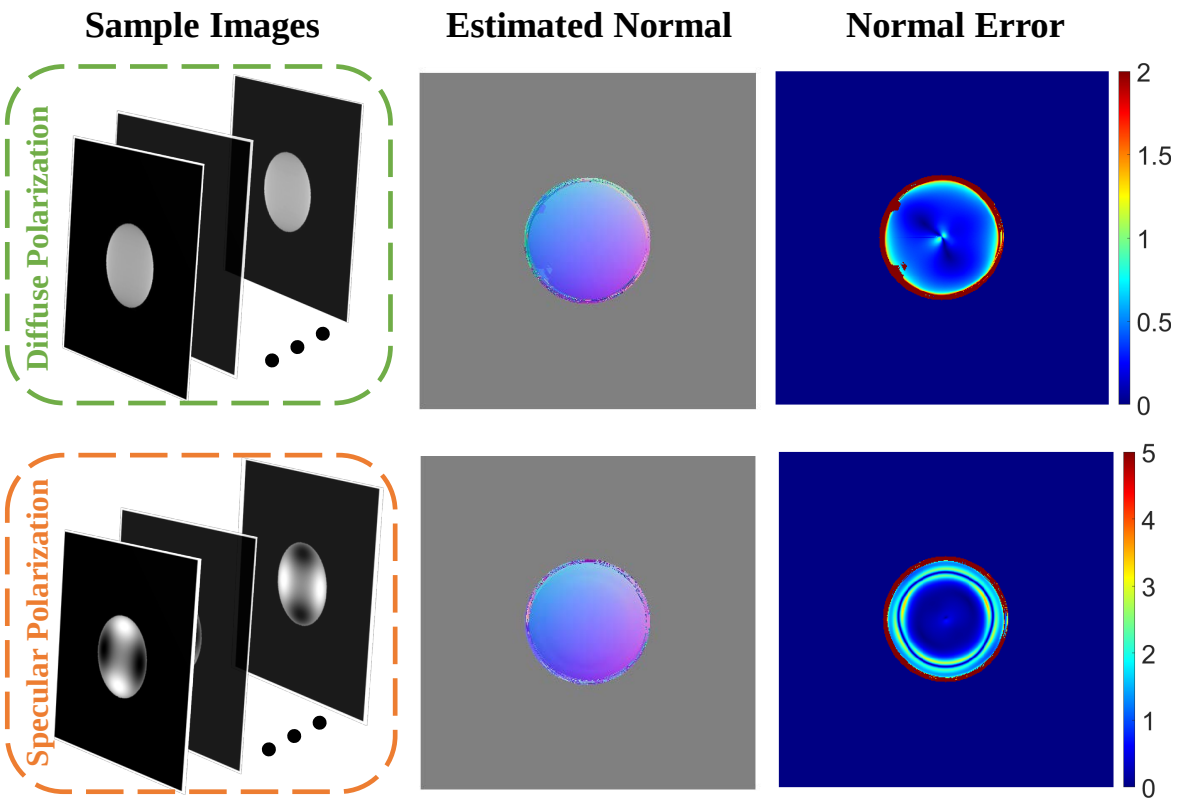


Figure 5.5: Qualitative synthetic results. The first column shows the input sample images. The second and third columns are our estimated normal maps and corresponding error maps.

5.4.3 Real Results

Next, We test our reconstruction algorithm on a real object with ceramic material, and the reconstruction result is shown in Fig. 5.6. We find that our reconstructed surface normal results do not contain ambiguous shape recovery in SfP. Since our method involves polarization information which can estimate the surface normal not only on Lambertian points but also non-Lambertian points, it provides satisfying reconstruction results for glossy objects. The bear scene also contains more challenging textureless surface. The results demonstrate that our technique is robust, effective and capable of handling extremely challenging specular and textureless objects with unknown refractive index and surface reflectance.

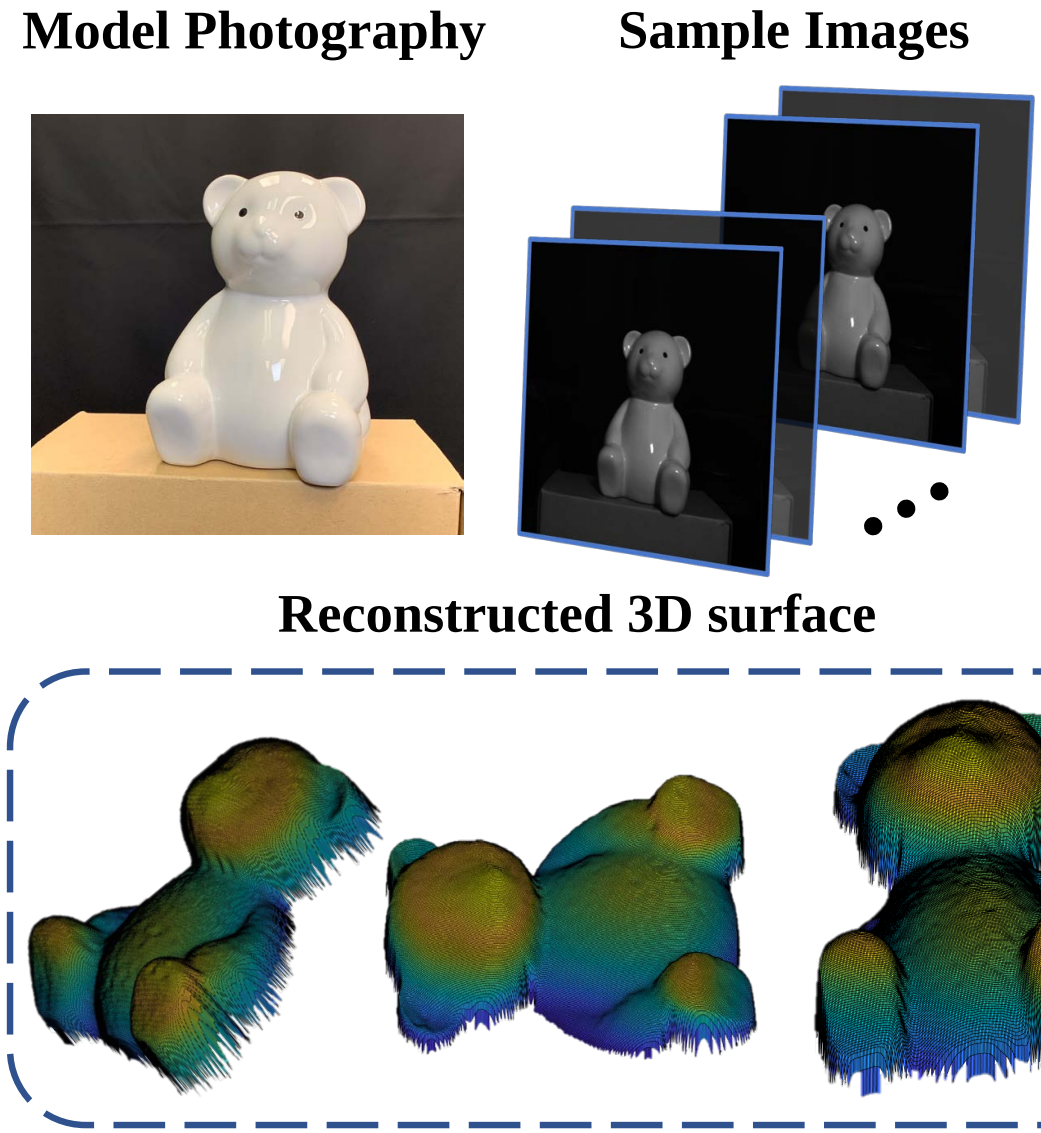


Figure 5.6: Shape recovery,the first row shows the visualization of the model and the sample images from our polarimetric plenoptic sampling. The shape recovery result is shown in the second row.

Chapter 6

CONCLUSIONS AND FUTURE WORK

In this dissertation, I have proposed three coded plenoptic sampling schemes. Each coded sampling information, such as non-centric sampling, spectral sampling and polarization sampling, can improve the efficiency of 3D reconstruction. On the acquisition front, I have developed the corresponding acquisition system for each of them. On the reconstruction front, I have introduced specific computational photography algorithms for 3D reconstruction via the coded sampling. This chapter first summarizes each plenoptic sampling scheme, and then introduces our future work.

6.1 Conclusions

6.1.1 R-XSlit Plenoptic Sampling

In Chapter 3, I have presented a new scheme to sample plenoptic function by using an XSlit camera. Different from previous pinhole based approaches that require translating cameras in 3D space, I fixed the XSlit camera in 3D space but rotated the slits. I have demonstrated that this acquisition scheme exhibits a significantly different sampling pattern of the plenoptic sampling. In particular, under this sampling pattern, any virtual perspective camera contains a minimal number of acquired samples. The acquired plenoptic sampling can be further used for effective 3D reconstruction (stereo matching and space carving) and for image-based rendering (new view synthesis and dynamic refocusing). I have also derived defocus blur kernels for R-XSlit plenoptic sampling and validated our theories through comprehensive experiments on both synthetic and real data.

On the requirement of mechanically rotating the slit to sample the plenoptic function, I admit that this is a limitation in this initial study, although our scheme has two major advantages compared with regular acquisition with pinhole camera. First, if we fix one slit and

rotate the other, we will acquire a plenoptic sampling which has the same ray sampling pattern as translating a pinhole camera along a line. However, rotating the lens/camera is much easier than translating the camera along a line. Second, in cases such as endoscopic imaging, it is very difficult to translate a camera. Our rotation scheme successfully overcomes this limitation.

6.1.2 Wavelength Coded Plenoptic Sampling

In Chapter 4, I have firstly presented the CWC plenoptic sampling scheme to recover specular surfaces. By using the MSS-Cam, I have proposed new consistency measurements to separate Lambertian points and non-Lambertian points, and then I have developed a robust algorithm to reconstruct shape and reflectance for surfaces with specularity that benefits of a wide range of vision and graphics tasks.

Second, I have developed a collaborative hybrid infrared sensing system based on our proposed IRWC plenoptic sampling scheme by combining computational imaging and illumination so as to achieve 3D face reconstruction under low-light and uncooperative conditions. Our method uses the special bright-eye effect of human eyes to facilitate 3D eye localization, which can be used to accurately determine the face pose and geometry. This will lead to a new class of face registration algorithms featuring 2D + 3D concepts. And our active hardware eye localization solution is more accurate, robust and fast. It also allows for fast face/head movement, common for unconstrained conditions.

6.1.3 Polarimetric Plenoptic Sampling

In Chapter 5, I have presented a novel polarimetric plenoptic sampling scheme and a framework for recovering specular and textureless objects. Traditional shape-from-polarization techniques suffer from the azimuth-zenith angle ambiguity. I have proved that our sampling scheme effectively resolves this ambiguity. I have derived a comprehensive theory to correlate the polarization radiance function with both surface normal and depth, and then I have derived a new differential formulation under our polarimetric plenoptic sampling for shape

reconstruction. Comprehensive experiments on synthetic and real data demonstrate our proposed technique is robust and effective for specular and textureless object reconstruction.

6.2 Future work

There are several exciting directions that I plan to explore in the future. For the R-XSlit plenoptic sampling, our immediate future work is to conduct experiments that individually rotate each slit to acquire the complete 4D plenoptic sampling. There are many interesting questions regarding the resulting light field including the ray density distribution when compared with the LF camera based on microlenslet array, its effects on refocusing quality (aliasing vs. blur kernel), its usefulness in depth inference, etc.

Moreover, the R-XSlit plenoptic sampling also reveals a previously overlooked property: a plenoptic sampling acquired by a multi-perspective camera is potentially better for rendering perspective images. This is illustrated in the ray density analysis in image-based rendering. On the contrary, the same argument can be made that a plenoptic sampling acquired by a perspective camera (e.g., a camera array) can better render a multi-perspective virtual view. We can interpret such a phenomenon in terms of ray geometry in 4D space as an image, perspective or multi-perspective, is a 2D planar cut (the General Linear Camera) in the ray space where ray samples can be viewed as intersections of the GLC plane with the sampling camera planes. In the future, I plan to study the corresponding theories and validate them through experiments using various plenoptic sampling acquisition solutions.

For other plenoptic sampling schemes (CWC plenoptic sampling, IRWC plenoptic sampling, and polarimetric plenoptic sampling), I acquired them by perspective camera arrays. Their reconstruction approaches can be potentially combined with the small baseline plenoptic sampling analysis (similar to [59]) to enhance geometric details to retrieve better initialization at the first glance. In fact, it will be highly desirable to develop a multi-scale solution to handle objects/scenes at different scales, which is also part of our immediate future work. For the infrared hybrid camera array, I can increase the number of infrared cameras and extend the near infrared flashing working range to recover high-resolution geometric

details from near infrared imaging. For the CWC and polarimetric plenoptic sampling acquisitions, instead of using the translation stage, I can build two camera arrays with spectral filters and polarizers respectively to reconstruct a dynamic scene.

My current approaches fail to use the semantic information. Recent advances on deep learning [71] exploit rich appearance data for modeling surface and can be integrated into our framework to pre-partition plenoptic sampling contents according to their material types. On the other hand, I intend to explore alternative surface reflectance models, including the ones obtained through deep learning, for more efficient geometry and material approximations.

BIBLIOGRAPHY

[1] Lytrolife in a different light. https://www.lytro.com/.

[2] Raytrix: 3d light field camera technology. http://www.raytrix.de/.

[3] Edward H Adelson, James R Bergen, et al. The plenoptic function and the elements of early vision. 1991.

[4] Amit Aides, Tamar Avraham, and Yoav Y Schechner. Multiscale ultrawide foveated video extrapolation. In *Computational Photography (ICCP), 2011 IEEE International Conference on*, pages 1–8. IEEE, 2011.

[5] Robert Anderson, Björn Stenger, and Roberto Cipolla. Color photometric stereo for multicolored surfaces. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2182–2189. IEEE, 2011.

[6] Gary A Atkinson and Edwin R Hancock. Multi-view surface reconstruction using polarization. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 309–316. IEEE, 2005.

[7] Gary A Atkinson and Edwin R Hancock. Recovery of surface orientation from diffuse polarization. *IEEE transactions on image processing*, 15(6):1653–1664, 2006.

[8] Gary A Atkinson and Edwin R Hancock. Shape estimation using polarization and shading from two views. *IEEE transactions on pattern analysis and machine intelligence*, 29(11), 2007.

[9] Gary A. Atkinson and Edwin R. Hancock. Shape estimation using polarization and shading from two views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29, 2007.

[10] Tamar Avraham and Yoav Y Schechner. Ultrawide foveated video extrapolation. *Selected Topics in Signal Processing, IEEE Journal of*, 5(2):321–334, 2011.

[11] Yuri Boykov and Gareth Funka-Lea. Graph cuts and efficient nd image segmentation. *IJCV*, 2006.

[12] Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *TPAMI*, 2004.

[13] Julie A Brefczynski and Edgar A DeYoe. A physiological correlate of the'spotlight'of visual attention. *Nature neuroscience*, 2(4):370–374, 1999.

[14] Jin-Xiang Chai, Xin Tong, Shing-Chow Chan, and Heung-Yeung Shum. Plenoptic sampling. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 2000.

[15] Can Chen, Haiting Lin, Zhan Yu, Sing Bing Kang, and Jingyi Yu. Light field stereo matching using bilateral statistics of surface cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1518–1525, 2014.

[16] Zhaopeng Cui, Jinwei Gu, Boxin Shi, Ping Tan, and Jan Kautz. Polarimetric multi-view stereo. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2017.

[17] James T Enns and Ronald A Rensink. Influence of scene-based properties on visual search. *Science*, 247(4943):721–723, 1990.

[18] Doron Feldman, Daphna Weinshall, et al. On the epipolar geometry of the crossed-slits projection. In *null*, page 988. IEEE, 2003.

[19] Graham Fyffe, Xueming Yu, and Paul Debevec. Single-shot photometric stereo by spectral multiplexing. In *Computational Photography (ICCP), 2011 IEEE International Conference on*, pages 1–6. IEEE, 2011.

[20] Steven J Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F Cohen. The lumigraph. In *SIGGRAPH*. ACM, 1996.

[21] Marcel Gutsche, Hendrik Schilling, Maximilian Diebold, and Christoph Garbe. Surface normal reconstruction from specular information in light field data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[22] Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. Effective face frontalization in unconstrained images. *arXiv preprint arXiv:1411.7964*, 2014.

[23] Carlos Hernández and George Vogiatzis. Self-calibrating a real-time monocular 3d facial capture system. In *Proceedings international symposium on 3D data processing, visualization and transmission (3DPVT)*, volume 2, 2010.

[24] Carlos Hernández, George Vogiatzis, Gabriel J Brostow, Bjorn Stenger, and Roberto Cipolla. Non-rigid photometric stereo with colored lights. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[25] Heiko Hirschmüller, Peter R Innocent, and Jon Garibaldi. Real-time correlation-based stereo vision with reduced border errors. *International Journal of Computer Vision*, 47(1-3):229–246, 2002.

[26] Berthold KP Horn. *Obtaining shape from shading information*. MIT press, 1989.

[27] Thomas E Hutchinson. Eye movement detector with improved calibration and speed, August 21 1990. US Patent 4,950,069.

[28] Cong Phuoc Huynh, Antonio Robles-Kelly, and Edwin Hancock. Shape and refractive index recovery from single-view polarisation images. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1229–1236. IEEE, 2010.

[29] Aaron Isaksen, Leonard McMillan, and Steven J Gortler. Dynamically reparameterized light fields. In *SIGGRAPH*. ACM, 2000.

[30] Martin Klaudiny and Adrian Hilton. High-detail 3d capture and non-sequential alignment of facial performance. In *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2012 Second International Conference on*, pages 17–24. IEEE, 2012.

[31] Kiriakos N Kutulakos and Steven M Seitz. A theory of shape by space carving. *IJCV*, 2000.

[32] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o (n) solution to the pnp problem. *International journal of computer vision*, 81(2):155–166, 2009.

[33] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 31–42. ACM, 1996.

[34] Nianyi Li, Jinwei Ye, Yu Ji, Haibin Ling, and Jingyi Yu. Saliency detection on light field. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[35] Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, and Paul Debevec. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Proceedings of the 18th Eurographics conference on Rendering Techniques*, pages 183–194. Eurographics Association, 2007.

[36] Laurence T Maloney. Evaluation of linear models of surface spectral reflectance with small numbers of parameters. *JOSA A*, 3(10):1673–1683, 1986.

[37] Daisuke Miyazaki, Masataka Kagesawa, and Katsushi Ikeuchi. Transparent surface modeling from a pair of polarization images. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (1):73–82, 2004.

[38] Daisuke Miyazaki, Robby T Tan, Kenji Hara, and Katsushi Ikeuchi. Polarization-based inverse rendering from a single view. In *null*, page 982. IEEE, 2003.

[39] Daisuke Miyazaki, Robby T. Tan, Kenji Hara, and Katsushi Ikeuchi. Polarization-based inverse rendering from a single view. In *ICCV*, 2003.

[40] Olivier Morel, Fabrice Meriaudeau, Christophe Stolz, and Patrick Gorria. Polarization imaging applied to 3d reconstruction of specular metallic surfaces. In *Machine Vision Applications in Industrial Inspection XIII*, volume 5679, pages 178–187. International Society for Optics and Photonics, 2005.

[41] MJ Morgan and RJ Watt. Mechanisms of interpolation in human spatial vision. *Nature*, 299(5883):553–555, 1982.

[42] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan. Light field photography with a hand-held plenoptic camera. *Computer Science Technical Report CSTR*, 2(11):1–11, 2005.

[43] Tomáš Pajdla. Geometry of two-slit camera. *Rapport Technique CTU-CMP-2002-02, Center for Machine Perception, Czech Technical University, Prague*, 2002.

[44] Tomáš Pajdla. Stereo with oblique cameras. *IJCV*, 2002.

[45] James F Parker Jr and Vita R West. Bioastronautics data book: Nasa sp-3006. *NASA Special Publication*, 3006, 1973.

[46] Jussi PS Parkkinen, J Hallikainen, and T Jaaskelainen. Characteristic spectra of munsell colors. *JOSA A*, 6(2):318–322, 1989.

[47] Stefan Rahmann and Nikos Canterakis. Reconstruction of specular surfaces using polarization imaging. In *null*, page 149. IEEE, 2001.

[48] Stefan Rahmann and Nikos Canterakis. Reconstruction of specular surfaces using polarization imaging. In *CVPR*, 2001.

[49] Steven M Seitz and Jiwon Kim. The space of all stereo images. *IJCV*, 2002.

[50] Steven A. Shafer. Using color to separate reflection components. *Color Research & Application*, 10(4):210–218, 1985.

[51] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3476–3483. IEEE, 2013.

[52] Michael W Tao, Jong-Chyi Su, Ting-Chun Wang, Jitendra Malik, and Ravi Ramamoorthi. Depth estimation and specular removal for glossy surfaces using point and line consistency with light-field cameras. *IEEE transactions on pattern analysis and machine intelligence*, 38(6):1155–1169, 2016.

[53] Shoji Tominaga and Norihiro Tanaka. Estimating reflection parameters from a single color image. *IEEE Computer Graphics and Applications*, 20(5):58–66, 2000.

[54] Stanford University. The (new) stanford light field archive, 2008. http://lightfield.stanford.edu/.

[55] Jan Van de Kraats and Dirk van Norren. Directional and nondirectional spectral reflection from the human fovea. *Journal of biomedical optics*, 13(2):024010–024010, 2008.

[56] Ashok Veeraraghavan, Ramesh Raskar, Amit Agrawal, Ankit Mohan, and Jack Tumblin. Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing. *TOG*, 2007.

[57] Daniel Vlasic, Pieter Peers, Ilya Baran, Paul Debevec, Jovan Popović, Szymon Rusinkiewicz, and Wojciech Matusik. Dynamic shape capture using multi-view photometric stereo. In *ACM Transactions on Graphics (TOG)*, volume 28, page 174. ACM, 2009.

[58] Ting-Chun Wang, Manmohan Chandraker, Alexei A Efros, and Ravi Ramamoorthi. Svbrdf-invariant shape and reflectance estimation from light-field cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5451–5459, 2016.

[59] Sven Wanner and Bastian Goldluecke. Globally consistent depth labeling of 4d light fields. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 41–48. IEEE, 2012.

[60] Lawrence B. Wolff. Polarization vision: a new sensory approach to image understanding. *Image and Vision Computing*, 15(2):81 – 93, 1997.

[61] Robert J Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1):191139, 1980.

[62] Robert J Woodham. Gradient and curvature from the photometric-stereo method, including local confidence estimation. *JOSA A*, 11(11):3050–3068, 1994.

[63] Jinwei Ye, Yu Ji, Wei Yang, and Jingyi Yu. Depth-of-field and coded aperture imaging on xslit lens. In *Computer Vision–ECCV 2014*, pages 753–766. Springer, 2014.

[64] Jinwei Ye, Yu Ji, and Jingyi Yu. Manhattan scene understanding via xslit imaging. In *CVPR*. IEEE, 2013.

[65] Jinwei Ye, Yu Ji, and Jingyi Yu. A rotational stereo model based on xslit imaging. In *ICCV*. IEEE, 2013.

[66] Jingyi Yu. *General linear cameras: theory and applications*. PhD thesis, Massachusetts Institute of Technology, 2005.

[67] Jingyi Yu and Leonard McMillan. General linear cameras. In *ECCV*. Springer, 2004.

[68] Jingyi Yu, Leonard McMillan, and Steven Gortler. Scam light field rendering. In *Computer Graphics and Applications, 2002. Proceedings. 10th Pacific Conference on*, pages 137–144. IEEE, 2002.

[69] Jingyi Yu, Leonard McMillan, and Steven Gortler. Surface camera (scam) light field rendering. *International Journal of Image and Graphics*, 4(04):605–625, 2004.

[70] Zhan Yu, Xinqing Guo, Haibing Ling, Andrew Lumsdaine, and Jingyi Yu. Line assisted light field triangulation and stereo matching. In *ICCV*. IEEE, 2013.

[71] Hang Zhang, Kristin J. Dana, and Ko Nishino. Friction from reflectance: Deep reflectance codes for predicting physical surface properties from one-shot in-field reflectance. *CoRR*, abs/1603.07998, 2016.

[72] Yingliang Zhang, Zhong Li, Wei Yang, Peihong Yu, Halting Lin, and Jingyi Yu. The light field 3d scanner. In *Computational Photography (ICCP), 2017 IEEE International Conference on*, pages 1–9. IEEE, 2017.

[73] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(11):1330–1334, November 2000.

[74] C Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High-quality video view interpolation using a layered representation. In *TOG*. ACM, 2004.

[75] Assaf Zomet, Doron Feldman, Shmuel Peleg, and Daphna Weinshall. Mosaicing new views: The crossed-slits projection. *TPAMI*, 2003.

Appendix A

PERMISSION LETTERS

The following are the permission letter from Mrs. Nianyi Li for using her publication in this dissertation and the permission letters from Mr. Yu Ji, Mr. Wei Yang, Mr. Haiting Lin, Mr. Yang Yang, Mr. Xinqing Guo, Mr. Zhong Li, and Mrs. Nianyi Li for using their facial images in this dissertation.

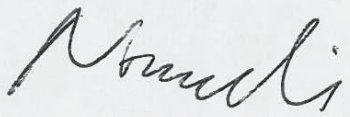
Sub: Permission Letter

To whom it may concern:

I'm writing this letter to give permission to my labmate Mingyuan Zhou to use our co-authored paper "Rotational Cross-Slit Light Field" published in Computer Vision and Pattern Recognition (CVPR) 2016 in his dissertation. I do not, however, give permission for any other use or for any re-disclosure of this information.

Yours Faithfully,

Nianyi Li



Sub: Permission Letter

To whom it may concern:

I'm writing this letter to give permission to my labmate Mingyuan Zhou to use my facial images in his paper, dissertation, and other forms of publications for research purposes only. I do not, however, give permission for any other use or for any re-disclosure of this information.

Yours Faithfully,

Yu Ji

Date: 09/01/2015

Sub: Permission Letter

To whom it may concern:

I'm writing this letter to give permission to my labmate Mingyuan Zhou to use my facial images in his paper, dissertation, and other forms of publications for research purposes only. I do not, however, give permission for any other use or for any re-disclosure of this information.

Yours Faithfully,

Wei Yang  Date: 09/01/2015

Sub: Permission Letter

To whom it may concern:

I'm writing this letter to give permission to my labmate Mingyuan Zhou to use my facial images in his paper, dissertation, and other forms of publications for research purposes only. I do not, however, give permission for any other use or for any re-disclosure of this information.

Yours Faithfully,

Haiting Lin Date: 09/01/2015

Sub: Permission Letter

To whom it may concern:

I'm writing this letter to give permission to my labmate Mingyuan Zhou to use my facial images in his paper, dissertation, and other forms of publications for research purposes only. I do not, however, give permission for any other use or for any re-disclosure of this information.

Yours Faithfully,

Yang Yang

Date: 09/01/2015

Sub: Permission Letter

To whom it may concern:

I'm writing this letter to give permission to my labmate Mingyuan Zhou to use my facial images in his paper, dissertation, and other forms of publications for research purposes only. I do not, however, give permission for any other use or for any re-disclosure of this information.

Yours Faithfully,

Xinqing Guo Date: 09/01/2015

Sub: Permission Letter

To whom it may concern:

I'm writing this letter to give permission to my labmate Mingyuan Zhou to use my facial images in his paper, dissertation, and other forms of publications for research purposes only. I do not, however, give permission for any other use or for any re-disclosure of this information.

Yours Faithfully,

Zhong Li Date: 09/01/2015

Sub: Permission Letter

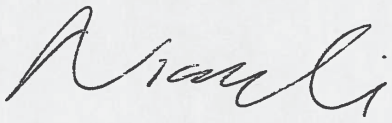
To whom it may concern:

I'm writing this letter to give permission to my labmate Mingyuan Zhou to use my facial images in his paper, dissertation, and other forms of publications for research purposes only. I do not, however, give permission for any other use or for any re-disclosure of this information.

Yours Faithfully,

Nianyi Li

Date: 09/01/2015



Copyright permission request

pubscopyright <copyright@osa.org> Mon, Feb 4, 2019 at 2:51 PM
To: Mingyuan Zhou <mzhou@udel.edu>, pubscopyright <copyright@osa.org>

Dear Mingyuan Zhou,

Thank you for contacting The Optical Society.

For the use of material from Mingyuan Zhou, Haiting Lin, S. Susan Young, and Jingyi Yu, "Hybrid sensing face detection and registration for low-light and unconstrained conditions," Appl. Opt. 57, 69-78 (2018):

> Because you are the author of the source paper from which you wish to reproduce material, OSA considers your requested use of its copyrighted materials to be permissible within the author rights granted in the Copyright Transfer Agreement submitted by the requester on acceptance for publication of his/her manuscript. If the entire article is being included, it is requested that the **Author Accepted Manuscript** (or preprint) version be the version included within the thesis and that a complete citation of the original material be included in any publication. This permission assumes that the material was not reproduced from another source when published in the original publication.
>
> The **Author Accepted Manuscript** version is the preprint version of the article that was accepted for publication but not yet prepared and/or formatted by The Optical Society or its vendors.

While your publisher should be able to provide additional guidance, OSA prefers the below citation formats:

> For citations in figure captions:
>
> [Reprinted/Adapted] with permission from ref [x], [Publisher]. (with full citation in reference list)
>
> For images without captions:
>
> Journal Vol. #, first page (year published) An example: Appl. Opt. 57, 69 (2018)

Please let me know if you have any questions.

Kind Regards,

Rebecca Robinson