

Algorithms for 3D Time-of-Flight Imaging

by

Jonathan Mei

Submitted to the Department of Electrical Engineering and Computer
Science

in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2013

© Jonathan Mei, MMXIII. All rights reserved.

The author hereby grants to MIT permission to reproduce and to
distribute publicly paper and electronic copies of this thesis document
in whole or in part in any medium now known or hereafter created.

Author
Department of Electrical Engineering and Computer Science
May 24, 2013

Certified by
Vivek K. Goyal
Research Scientist, Research Laboratory of Electronics
Thesis Supervisor

Accepted by
Prof. Dennis M. Freeman
Chairman, Masters of Engineering Thesis Committee

Algorithms for 3D Time-of-Flight Imaging

by

Jonathan Mei

Submitted to the Department of Electrical Engineering and Computer Science
on May 24, 2013, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering

Abstract

This thesis describes the design and implementation of two novel frameworks and processing schemes for 3D imaging based on time-of-flight (TOF) principles. The first is a low power, low hardware complexity technique based on parametric signal processing for orienting and localizing simple planar scenes. The second is an improved method for simultaneously performing phase unwrapping and denoising for sinusoidal amplitude modulated continuous wave ToF cameras using multiple frequencies.

The first application uses several unfocused photodetectors with high time resolution to estimate information about features in the scene. Because the time profiles of the responses for each sensor are parametric in nature, the recovery algorithm uses finite rate of innovation (FRI) methods to estimate signal parameters. The signal parameters are then used to recover the scene features.

The second application uses a generalized approximate message passing (GAMP) framework to incorporate both accurate probabilistic modeling for the measurement process and underlying scene depth map sparsity to accurately extend the unambiguous depth range of the camera. This joint processing results in improved performance over separate unwrapping and denoising steps.

Thesis Supervisor: Vivek K. Goyal

Title: Research Scientist, Research Laboratory of Electronics

Acknowledgments

I would like to thank my wonderful mom and dad for their unwavering love and support throughout my life, especially in the difficult times during my studies, and Vincent for being such a good friend and little brother. I want to thank Vivek K Goyal for his insight and guidance through this significant and enriching academic experience, as well as for offering invaluable perspective and advice in helping me make important decisions continuing forward. I also want to thank Ahmed Kirmani for acting as a mentor in my transition into the STIR group and the rest of the group for being so welcoming and making me feel at home in the office. I would like to thank my friends for putting up with my ramblings and everyday antics. Finally, I want to thank Mariya Toneva for her constant encouragement even when I was doubtful of myself, helping me to keep the bigger picture in mind.

This material is based upon work supported in part by the National Science Foundation under Grant No. 1101147 and the HP Labs Innovation Research Program.

Contents

1	Introduction	11
1.1	Time-of-Flight Principle	12
1.2	Physical modeling	13
1.3	Signal representations	14
1.3.1	Parametric Signals	14
1.3.2	Sparse Signals	14
2	Planar Scene Feature Recovery (PSFeaR)	17
2.1	Device and Scene Setup	18
2.1.1	Single Plane Scene Impulse Response	19
2.1.2	Sampling the Sensor Response	20
2.1.3	Forward Modeling for the Sensor Response	21
2.1.4	Multiple Plane Scene Impulse Response	23
2.2	Estimating Plane Pose	23
2.3	PSFeaR Algorithm	25
2.3.1	Parametric Sampling Step	26
2.3.2	Recovery Step	28
2.3.3	Scene Containing Multiple Planes	28
2.4	Simulations and Discussion	29
2.4.1	Performance of Parametric Sampling Step	29
2.4.2	PSFeaR Performance	32
2.5	Conclusion	35

3	Simultaneous Phase Unwrapping and Denoising (SPUD)	37
3.1	Time-of-Flight Imaging	37
3.2	Operation of TOF Cameras	38
3.2.1	Homodyne Measurement and Conventional Phase Estimation	38
3.2.2	Statistical Measurement	40
3.2.3	Phase Unwrapping	41
3.3	Unwrapping and Denoising TOF Measurements	42
3.3.1	Transform Domain Signal Modeling	42
3.3.2	GAMP-Based Unwrapping and Denoising	43
3.3.3	Approximating the Output Nonlinear Step	45
3.3.4	Approximating the Input Nonlinear Step	49
3.3.5	Computation	49
3.4	Simulations and Discussion	50
4	Conclusion	55
4.1	PSFeaR	55
4.2	SPUD	56

List of Figures

2-1	Device and example scene of interest composed of a single infinite plane	18
2-2	Device and example scene of interest composed of a two semi-infinite planes	19
2-3	Signal flow diagram for signal acquisition at Sensor k	20
2-4	Typical single infinite plane (a) scene impulse response and (b) sensor response	22
2-5	Scene impulse responses and sensor responses from two planes being imaged internally	24
2-6	Approximate time derivative of scene impulse response	27
2-7	Variance of estimates of onset times τ for varying model order with fixed approximation order (0th order box fit)	30
2-8	Variance of estimates of onset times τ for 0th order (box) and 1st order (linear) approximations for fixed model order $L = 6$	31
2-9	Average normalized bias over typical range of scenes	32
2-10	Estimates of scene impulse responses using the full forward modeling based recovery.	33
2-11	Time derivatives and estimated Dirac locations of scene impulse response for simulated scene containing (a) one plane and (b) two planes	33
2-12	Reconstruction of scene impulse response from estimated Diracs for same simulated scenes as in Figure 2-11 of (a) one plane and (b) two planes.	34
2-13	Effect of noise on error in estimation of plane parameters for each of two planes. Each noise level was simulated 50 times	34

3-1	Signal processing abstraction of TOF camera	39
3-2	Forward acquisition model for the TOF camera measurements at two modulation frequencies.	43
3-3	One iteration of our GAMP algorithm for unwrapping and denoising. This diagram shows the updates of the estimates for $\hat{\mathbf{z}}(t)$ and $\hat{\mathbf{x}}(t)$, represented by dark green nodes. The means of the estimates $\hat{\boldsymbol{\mu}}_z(t)$ and $\hat{\boldsymbol{\mu}}_x(t)$ and the variances $\hat{\boldsymbol{\sigma}}_z(t)$ and $\hat{\boldsymbol{\sigma}}_x(t)$ are updated concurrently at each step	47
3-4	Differences between sum of Gaussian approximation and exponential of cosine functions for low frequency, high frequency, and two frequencies at $k = 0.4$	48
3-5	Errors of sum of Gaussian approximation for exponential of cosine likelihood functions for (3-5a) low frequency, (3-5b) high frequency, and (3-5c) two frequencies over a range of k values	48
3-6	Ground truth and estimates for “Tsukuba 450” scene computed from (3.9) using single modulation frequencies at SNR = 10 dB.	51
3-7	Reconstructed depth maps and their root mean squared errors. The proposed GAMP method provides a 0.5 dB improvement relative to the median filtered MLE and more than 5 dB relative to the pointwise MLE. All images produced at SNR = 10 dB.	51
3-8	Pixelwise standard deviations over 50 simulations at 10 dB SNR . . .	52
3-9	Root mean squared error comparison for several methods across an SNR range for “Tsukuba 450” image.	52
3-10	Comparison of filter lengths on “Tsukuba 450” image across an SNR range.	53

Chapter 1

Introduction

Imaging is the process of encoding physical information about a scene of interest to produce an image. In traditional optical imaging, the physical information captured describes color and light intensities reflected by objects in a scene. By contrast, in 3D imaging, the goal is to obtain spatial information about a scene.

In general, imaging can be accomplished by passive and active means. Passive imaging schemes utilize ambient light from the environment as illumination of the scene, while active imaging devices produce their own illumination and measure the response of the scene to the illumination.

3D imaging is a growing area of interest, as applications in robotics, security, health care, gaming, etc. increasingly incorporate 3D information. A variety of qualities are desirable for a 3D imaging system, including accuracy, robustness to ambient light, portability, and high frame rates. As a result, much research in 3D imaging technology focuses on noise performance, size, power, and processing speed.

The two new imaging methods introduced in this thesis, Planar Scene Feature Recovery (PSFeaR) and Simultaneous Phase Unwrapping and Denoising (SPUD), both are active 3D imaging technologies that make use of time-of-flight (TOF) principles and exploit physical modeling and sparsity to make improvements in these challenge areas. The remainder of this thesis will be organized into two major parts describing PSFeaR and SPUD as follows:

The rest of the introduction in Chapter 1 will lay out the fundamental concepts

and themes of TOF, physical modeling, and signal representation that underlie both imaging frameworks.

Chapter 2 describes the PSFeaR architecture and processing, first giving context for the diffuse imaging problem and describing the advantages and disadvantages of proposed framework in relation to current art. It presents a new imaging device and intended scenes of interest and then outlines the sampling techniques and estimation algorithms used to recover the scene features. Finally, the chapter evaluates the simulated performance of this framework for estimating simple scenes under various noise conditions.

Chapter 3 describes the SPUD algorithm and the operation of the TOF cameras it is intended for. It outlines the operating principles of an amplitude modulated continuous wave (AMCW) TOF camera and the appropriate forward modeling for its measurements. Then, it describes the generalized approximate message passing (GAMP) framework and the specific details of its implementation for solving the unwrapping and denoising problem. The chapter ends with simulations of the SPUD algorithm and comparisons of its performance to traditional separate processing methods.

Chapter 4 concludes the thesis with a discussion of possible future directions for the two frameworks

1.1 Time-of-Flight Principle

Time-of-flight cameras measure distance by relating the finite speed of light to the time traveled by light between the scene and the imaging device as $d = ct$. TOF systems are active imaging systems and can be categorized further into two groups, pulsed and continuous wave (CW), based on the illumination signal they produce.

Pulsed TOF systems produce a single pulse within one repetition period and measure a response in the period. The pulse shape is typically much narrower in time than the repetition period. While conventional systems may integrate the return signal and measure the time at which the integrated signal crosses a threshold as the approximate time-of-flight, PSFeaR will use parametric signal processing to measure

the time-of-flight. Continuous-wave ToF systems produce a periodic signal, such as a sinusoid, and measure the phase shift between the received signal and the original signal. This measurement of phase shift can be made by performing a cross correlation between the two signals. The conventional CW TOF architecture and processing will be presented with more detail in Chapter 3 when introducing SPUD.

1.2 Physical modeling

The operation and performance of optical imaging systems is in part determined by the fundamental physics governing how the intensity of light changes as it travels and how this intensity is measured electronically.

Diffuse or unfocused light from a point source experiences decreasing intensity as it radiates outward, exhibiting behavior described by the inverse square law. The intensity of a light signal measured a distance r from the source has intensity proportional to the inverse of distance squared ($\propto 1/r^2$). This radial falloff effect dictates the quality of the measurements corresponding to further points in a scene, especially in the presence of ambient light.

Measurement of light is achieved in semiconductors through a process of converting photons to electrons. This conversion process is probabilistic, and causes what is known as shot noise. The charge generation in photodetectors is well characterized as a time inhomogeneous Poisson process with rate λ proportional to the intensity of light [8], with discrete distribution

$$p(k) = \frac{e^{-\lambda} \lambda^k}{k!}, k \in \mathbb{N}_0$$

where k is a realization of a random variable denoting the number of electrons generated at some time instance and \mathbb{N}_0 denotes the set of non-negative integers. While many traditional processing methods assume this measurement is roughly described by a Gaussian random variable with fixed variance, the true nature of the measurement process is signal dependent.

PSFeaR and SPUD incorporate these two physical phenomena into their models to achieve improved performance.

1.3 Signal representations

As technology advances the ability to sample time series at a faster rate or capture more pixels in a fixed-size image, the amount of information used to represent these 1D and 2D signals increases. However, many interesting signals and images have certain properties that allow them to be represented or closely approximated in clever ways.

1.3.1 Parametric Signals

Many signals can be characterized by a finite set of parameters describing specific information about the signal. For example, a signal of the form $f(x) = ax^2 + bx + c$ may be parameterized by the triple of coefficients (a, b, c) for each power of x . An alternate parameterization representing different information about the signal could be the triple (d, h, r) for the representation $f(x) = d(x - h)^2 + r$. In either case, certain information about the signal is captured in the parameters, and the signal belongs to a parametric family.

In the PSFeaR framework, the parametric nature of the signals is exploited in the processing method to allow lower sampling rate and higher-resolution for recovery than classical sampling methods. This direct recovery of parameters also allows for faster estimation of the 3D scene than recovery from using a full forward signal model.

1.3.2 Sparse Signals

A signal may be said to have a sparse representation if its coefficients are near zero in an appropriately chosen transform domain. Images of natural scenes have been shown to have sparse representations in the discrete wavelet basis [14]. This sparsity also applies to depth maps, which tend to have wavelet coefficients that are approximately Laplacian in distribution, with density

$$p(x) = \frac{1}{2q} e^{-|x/q|} \quad (1.1)$$

for some parameter q .

In the SPUD algorithm, knowledge of signal sparsity is used to achieve denoising in the depth map produced.

Chapter 2

Planar Scene Feature Recovery (PSFeaR)

Planar Scene Feature Recovery (PSFeaR) is a new imaging architecture, in both measurement and processing, for directly sensing 3D structure of scenes composed of planar components. The measurement device has compact size and uses standard inexpensive hardware components. The algorithm employs the time-of-flight (TOF) principle and parametric signal processing techniques to achieve significantly lower computational complexity than computer vision methods based on stereo disparity or structured light [16] and methods relying on forward simulation.

Practical applications of interest, such as generating physically realistic rendered augmentations or 3D indoor localization and mapping, rely on only few specific scene features such as plane orientation and location. Current image processing and computer vision pipelines operate by obtaining a full 2D or 3D image of the scene, detecting objects, then estimating object parameters. While useful for general scenes, the pipeline requires significant acquisition and computation resources. The PSFeaR architecture obviates the need for these expensive requirements for simple planar scenes through the use of parametric modeling of scene impulse responses.

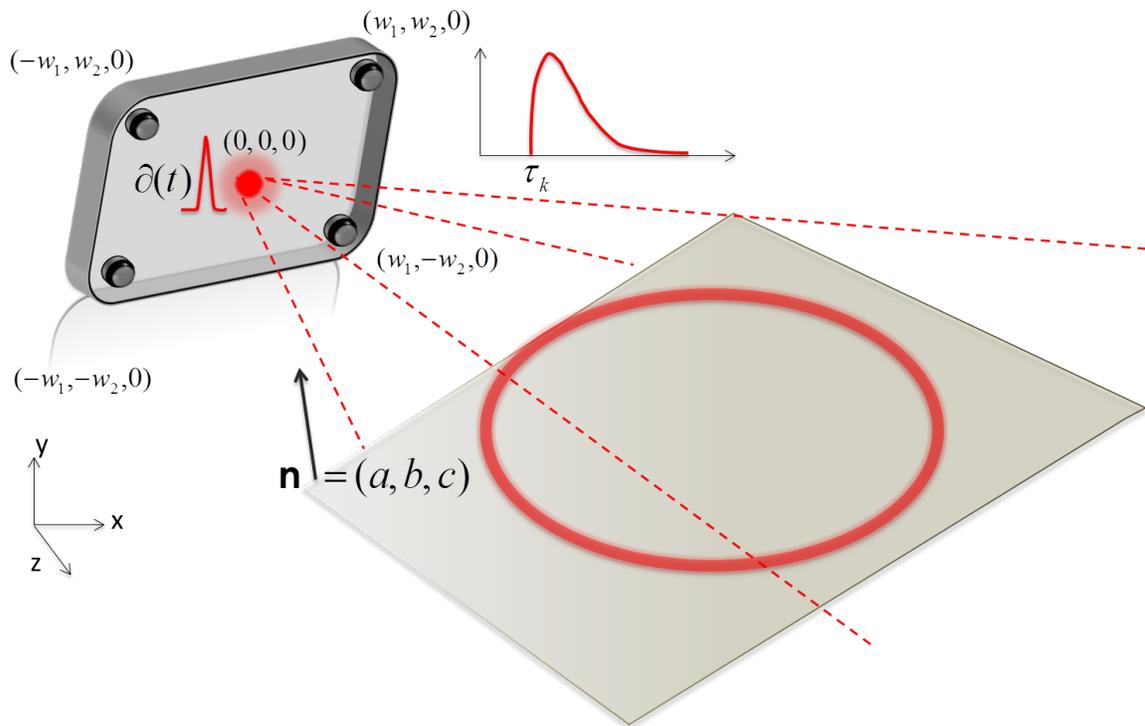


Figure 2-1: Device and example scene of interest composed of a single infinite plane

2.1 Device and Scene Setup

Consider the device and scene of interest shown in Figure 2-1, corresponding to the practical scenario of physically accurate rendering for augmented reality applications. The processing framework will ultimately be demonstrated for recovering the orientation and location of one or two planes in the camera field of view.

The proposed imaging architecture comprises of a single intensity modulated light source that illuminates the scene with a T -periodic signal $s(t)$ and 4 time-resolved sensors. The intensity of reflected light detected at Sensor k is $r_k(t)$, $k = 1, \dots, 4$. The light source and detectors are synchronized to a common time origin. It is also assumed that the illumination period is large enough to avoid aliasing and distance ambiguity [13]. To derive the scene impulse response, we let $s(t) = \delta(t)$.

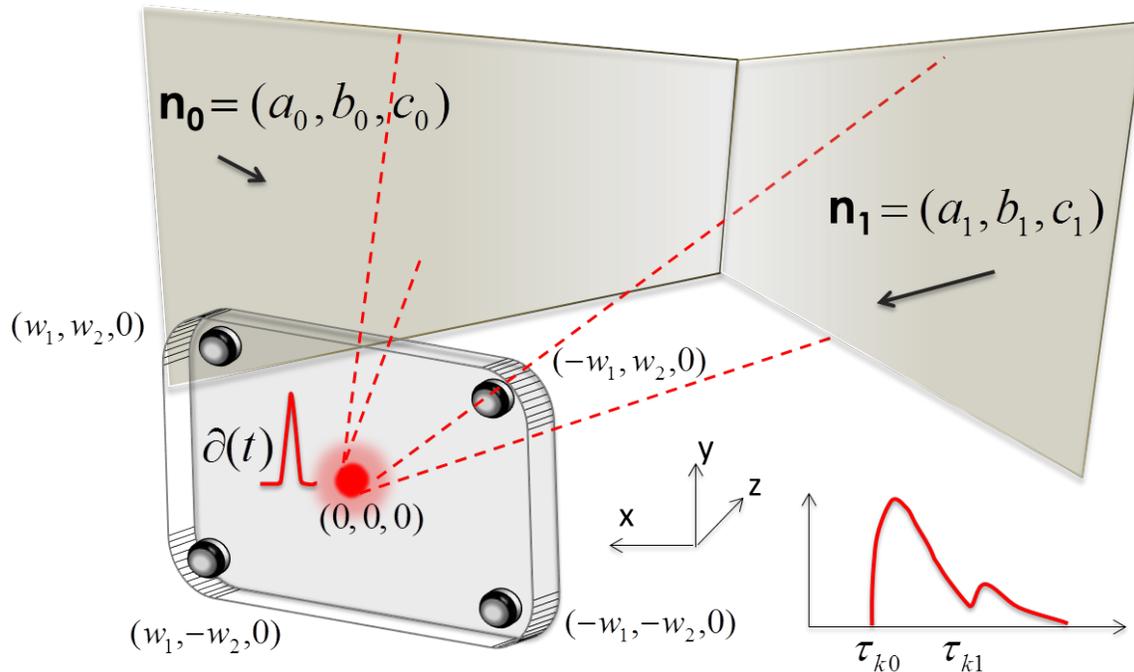


Figure 2-2: Device and example scene of interest composed of a two semi-infinite planes

2.1.1 Single Plane Scene Impulse Response

Consider a scene comprising a single infinite plane occupying the entire field of view (FOV). Following [6], let $\tilde{\mathbf{x}} = (\tilde{x}_1, \tilde{x}_2) \in \mathcal{R}^2$ be a point on the scene plane corresponding to point $\mathbf{x} = R(\tilde{\mathbf{x}}) = (R_1(\tilde{x}_1, \tilde{x}_2), R_2(\tilde{x}_1, \tilde{x}_2), R_3(\tilde{x}_1, \tilde{x}_2)) \in \mathcal{R}^3$ in the previously defined 3D space oriented relative to the device, let $d^{(s)}(\mathbf{x})$ denote the distance from the illumination source at $(0, 0, 0)$ to \mathbf{x} , and let $d_k^{(r)}(\mathbf{x})$ denote the distance from \mathbf{x} to Sensor k . Then $d_k^{(t)}(\mathbf{x}) = d^{(s)}(\mathbf{x}) + d_k^{(r)}(\mathbf{x})$ is the total distance traveled by the light contribution from \mathbf{x} . This contribution is attenuated by the reflectance $f(\mathbf{x})$, square-law radial fall-off, and $\cos_+(\theta(\mathbf{x})) = (\cos(\theta(\mathbf{x})) + |\cos(\theta(\mathbf{x}))|)/2$, for $0 \leq \theta(\mathbf{x}) \leq \pi$ to account for the illumination beam angle intensity profile, where $\theta(\mathbf{x})$ is the angle between \mathbf{x} and the vector $\langle 0, 0, 1 \rangle$ normal to the device. Here, the imaging direction is chosen to be in the positive z -direction so that the sensor's entire FOV lies in the halfspace where $z > 0$. The $\cos_+(\theta(\mathbf{x}))$ models stronger illumination of scene points more directly in front of the device than for those more lateral to the device.

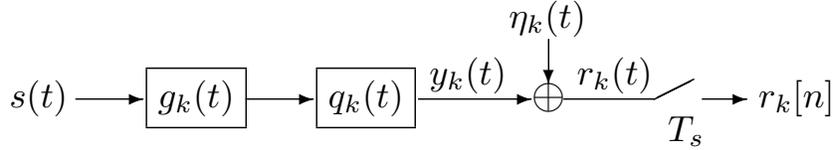


Figure 2-3: Signal flow diagram for signal acquisition at Sensor k .

Using $s(t) = \delta(t)$, the amplitude contribution from point \mathbf{x} is the light signal $a_k(\mathbf{x}) f(\mathbf{x}) \delta(t - d_k^{(t)}(\mathbf{x})/c)$ where

$$a_k(\mathbf{x}) = \cos_+(\theta(\mathbf{x})) / \left(d^{(s)}(\mathbf{x}) d_k^{(r)}(\mathbf{x}) \right)^2. \quad (2.1)$$

Combining contributions over the plane, the total light incident at Sensor k at time t is

$$g_k(t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} a_k(\mathbf{x}) f(\mathbf{x}) \delta(t - d_k^{(t)}(\mathbf{x})/c) dx_1 dx_2. \quad (2.2)$$

The intensity $g_k(t)$ thus contains the contour integrals over the object surface where the contours are ellipses. For $f(\mathbf{x}) > 0$, it can be seen that $g_k(t)$ is zero until a certain onset time τ_k . For constant reflectivity $f(\mathbf{x}) = 1$, $g_k(t)$ can be approximately represented in a parametric form and can also be well approximated by a polynomial spline of degree at most 1. This polynomial approximation will form the basis for the plane localization and orientation algorithm. For the remainder of the analysis in this thesis, it is assumed that the reflectivity $f(\mathbf{x}) = \alpha$ is constant but unknown.

2.1.2 Sampling the Sensor Response

An implementable digital system requires sampling at the detectors. Furthermore, a practical detector has an impulse response $q_k(t)$, and the Dirac impulse used in the above analysis is an abstraction that cannot be realized in practice. Using the fact that light transport is linear and time invariant, we accurately represent the signal acquisition pipeline at Sensor k using the flow diagram in Figure 2-3.

For a photon-counting detector, the signal can be modeled as a Poisson process

with rate:

$$y_k(t) = g_k(t) * q_k(t) * s(t)$$

Assuming operation at high light levels, digital samples can be well modeled as:

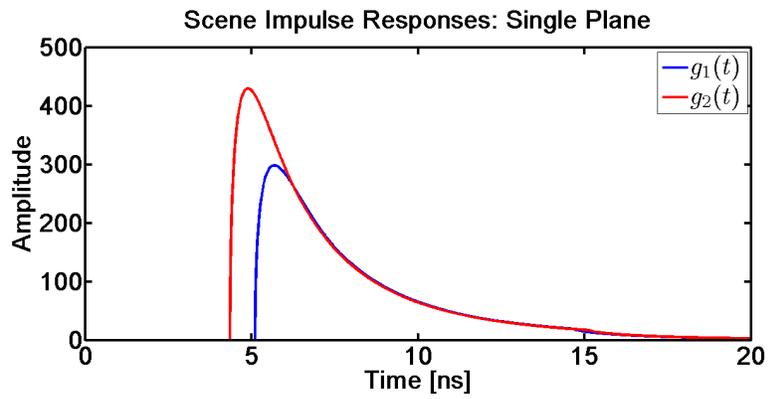
$$r_k[n] \approx y_k(t)|_{t=nT_s}, \quad n = 1, \dots, N,$$

where N such samples are acquired in each illumination period with a sampling period $T_s = T/N$. Samples $y_k[n] = y_k(t)|_{t=nT_s} = E\{r_k[n]\}$ for $n = 1, \dots, N$ can be calculated accurately given the exact plane orientation and the sensor coordinates. For the remainder of this thesis, it is assumed that the 4 detectors have identical responses $q_k(t) = q(t)$. Then the continuous ideal scene response at the detector $y_k(t) = g_k(t) * h(t)$ where $h(t) = q(t) * s(t)$ represents the combined pulse shape and sensor impulse response. Figure 2-4 shows a typical scene impulse response and sensor response from a scene composed of one infinite plane.

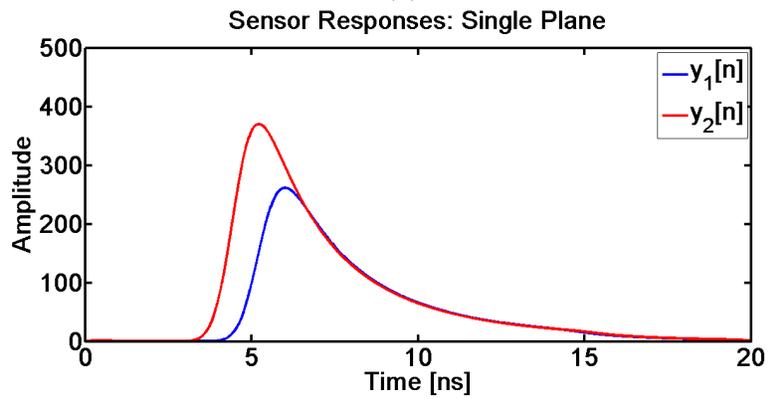
2.1.3 Forward Modeling for the Sensor Response

The scene response at the sensor is the convolution of the continuous scene impulse response $g_k(t)$ with the combined pulse shape and sensor impulse response $h(t)$. Assuming a known combined pulse shape and sensor impulse response $h(t)$, the scene impulse response $g_k(t)$ needs to be computed in order to find the scene response samples $r_k[n]$. The scene impulse response has been expressed as a contour integral in two variables on the plane, as in Equation 2.2. Because the contours are ellipses, the coordinate transformation R^{-1} from the original 3D Cartesian system to a 2D polar coordinate system confined to the plane can be applied to simplify the representation. In addition, the elliptical contour integral has no closed form solution even for a simple constant reflectivity function such as $f(x) = \alpha$, so a numerical integration is required to simulate impulse response values.

Since the closed form expression for $g_k(t)$ cannot be produced, samples of $g_k(t)$ can



(a)



(b)

Figure 2-4: Typical single infinite plane (a) scene impulse response and (b) sensor response

instead be numerically computed at given time instants as previously mentioned. To approximate a continuous time convolution, the oversampled $g_k^{(os)}[n] = g_k(t)|_{t=nT_s/M}$ can be convolved with oversampled $h^{(os)}[n] = h(t)|_{t=nT_s/M}$, where M is the oversampling rate. This yields $y_k^{(os)}[n] = h^{(os)}[n] * g_k^{(os)}[n]$, which can then be downsampled $\hat{y}[n] = y^{(os)}[Mn]$ to produce an accurate estimate for $\hat{y}[n] \approx y[n]$ at the appropriate sampling rate for the sensor.

2.1.4 Multiple Plane Scene Impulse Response

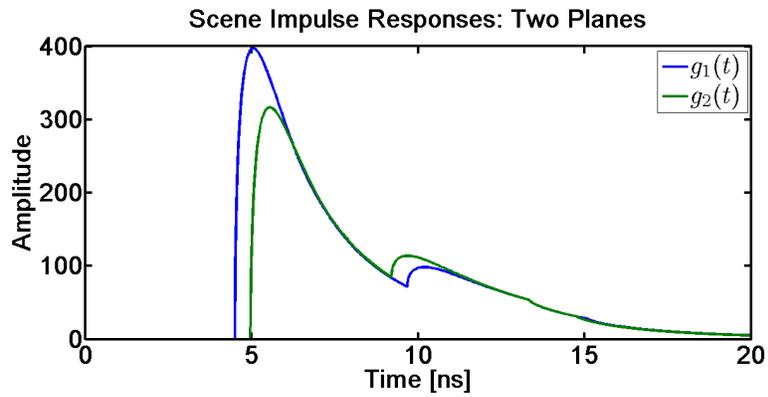
Consider a scene in which two planes intersect in the FOV of the device, corresponding to localizing the device with respect to the interior corner of a room. The impulse response can be calculated similarly to Eq. 2.2 describing the single plane case,

$$g_k(t) = \sum_{i=1}^2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} a_{ki}(\mathbf{x}) f_i(\mathbf{x}) \delta(t - d_{ki}^{(t)}(\mathbf{x})/c) dx_{1i} dx_{2i}. \quad (2.3)$$

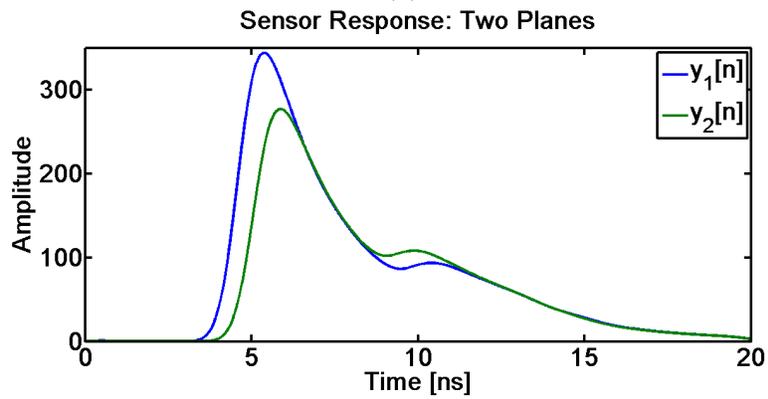
The impulse response from the two planes is initially identical to the sum of the individual plane responses. However, the response from the occluded semi-infinite portions of each plane results in an additional negative contribution, as shown in a typical two plane scene impulse response in Figure 2-5. In modeling the response, this occlusion can be simulated by setting $f_i(\mathbf{x}) = 0$ for points \mathbf{x} beyond the intersection of the planes. This computation can similarly be extended to 3 or more planes as well.

2.2 Estimating Plane Pose

With the ability to perform full forward modeling, it is now possible to estimate the plane pose from samples of the scene response. A plane $P(\mathbf{n})$ that does not intersect the origin can be parameterized by the 3D coordinates of the point $\mathbf{n} = (a, b, c)$ on the plane that lies closest to the origin. This plane is alternatively described by the equation $\mathbf{n} \cdot \mathbf{x} = \mathbf{n} \cdot \mathbf{n}$, which uniquely determines the normal of the plane as well as its location. For every such plane $P(\mathbf{n})$, samples of expected noiseless response at



(a)



(b)

Figure 2-5: Scene impulse responses and sensor responses from two planes being imaged internally

Sensor k , $y_k[n; a, b, c]$, can be computed. Then the estimated values for (a, b, c) can be posed as an optimization problem to find the $\hat{y}_k[n] = y_k[n; \hat{a}, \hat{b}, \hat{c}]$ with the minimum error relative to the collected noisy samples $r_k[n]$:

$$(\hat{a}, \hat{b}, \hat{c}) = \underset{(a,b,c)}{\operatorname{argmin}} \sum_k \sum_n (r_k[n] - y_k[n; a, b, c])^2 \quad (2.4)$$

Note that this recovery directly captures scene features without requiring acquisition of a complete 2D or 3D image. However, this formulation involves full forward simulation as an intermediate step, and as a result the objective function expressed in this form is difficult to characterize in terms of the parameters of interest (a, b, c) . Moreover, because there is no closed form expression for the impulse response of a scene, the accurate simulation of scene responses is an expensive computation. When dealing with multiple planes, the complexity becomes prohibitive. Other approaches are necessary in order for the computation to be tractable and for the imaging framework to be practical.

2.3 PSFeaR Algorithm

The approach that the PSFeaR framework takes to reduce the complexity of the computation is to incorporate parametric sampling. Once the important signal parameters are recovered, the scene features can be estimated from those parameters. This greatly reduces the computational cost of the estimation of plane orientations from the collection of time samples. This section will first outline the algorithm for a scene consisting of a single plane and then examine the recovery of a scene composed of two intersecting planes. In both cases, the framework applies a two-step process to the feature acquisition problem:

1. Use parametric deconvolution to estimate scene impulse responses $g_k(t)$ and onset times τ_k from samples $r_k[n]$.
2. Use the set of estimated signal parameters (onset times τ_k) to recover scene features.

2.3.1 Parametric Sampling Step

The central pillar of classical sampling methods is the Shannon sampling theorem, which dictates that perfect reconstruction of a continuous bandlimited signal from noiseless samples can be achieved with a sampling rate of twice the bandlimit, $\omega_s > 2\omega_{max}$. This perfect reconstruction is achieved through interpolation using a bandlimited *sinc* kernel. On the other hand, parametric sampling techniques utilize the fact that a signal to be sampled belongs to a parametric family to directly recover those specific parameters describing the continuous signal. Parametric methods may require lower sampling rates compared to traditional sampling and processing methods.

The Finite Rate of Innovation (FRI) method used in the PSFeaR framework is a parametric technique and in this case is used to accurately estimate onset times τ_k of the impulse response $g_k(t)$. FRI sampling allows recovery of a time signal composed of a stream of Dirac deltas blurred by a sampling kernel using fewer samples than dictated by Shannon theory in a tractable computation [17]. Similarly, FRI techniques also allow recovery of piecewise polynomial functions.

Consider a signal composed of a periodic stream Diracs of the form

$$x(t) = \sum_{\ell} a_{\ell} \sum_n \delta(t - t_{\ell} - nT)$$

By using Poisson's summation formula, we can rewrite

$$\begin{aligned} x(t) &= \sum_m \frac{1}{T} \left(\sum_{\ell} a_{\ell} e^{-i(2\pi m t_{\ell}/T)} \right) e^{i(2\pi m t/T)} \\ &= \sum_m X[m] e^{i(2\pi m t/T)} \end{aligned}$$

where $X[m]$ are the discrete Fourier coefficients of $x(t)$.

The time locations of the Diracs are directly related to the frequency of the complex exponentials that compose the discrete Fourier coefficients $X[m]$. The FRI technique uses spectral estimation methods, such as annihilating filter or matrix pencil, on $2L + 1$ of these discrete Fourier coefficients to efficiently find up to L frequencies

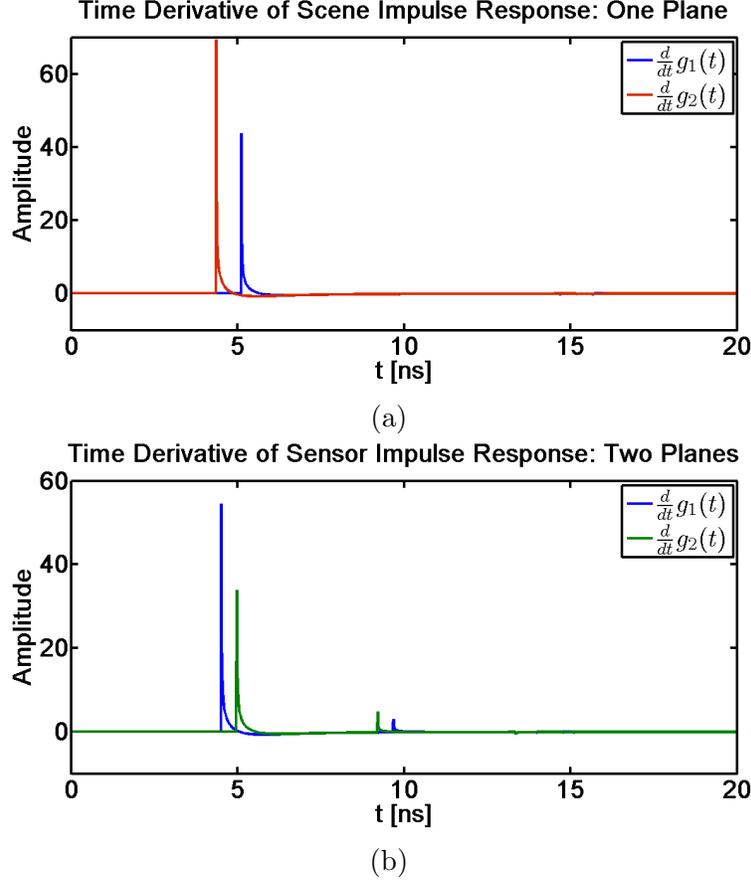


Figure 2-6: Approximate time derivative of scene impulse response

in the Fourier coefficients and thus up to L time locations of the Diracs. Once the time locations are found, the amplitudes can be recovered through solving a linear system

$$\begin{bmatrix} X[0] \\ X[1] \\ \vdots \\ X[L-1] \end{bmatrix} = \frac{1}{T} \begin{bmatrix} 1 & 1 & \dots & 1 \\ z_0 & z_1 & \dots & z_{L-1} \\ \vdots & \vdots & \ddots & \vdots \\ z_0^{L-1} & z_1^{L-1} & \dots & z_{L-1}^{L-1} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_{L-1} \end{bmatrix}$$

where $z_l = e^{-i(2\pi t_l/T)}$.

A noiseless sensor can only observe samples $y_k[n]$, but from samples $Y_k[m] = G_k[m]H[m]$ we can recover $G_k[m]$ with knowledge of $H[m]$. Although the measurements from the detector are not noiseless, the samples $r_k[n]$ can be used as an estimate for the noiseless values $y_k[n]$, giving $R_k[m]/H[m] \approx Y_k[m]/H[m] = G_k[m]$. With ap-

appropriate denoising techniques, the spectral estimation described previously can be performed well.

For recovery of onset times of the scene impulse responses $g_k(t)$, observing that the time derivatives $\frac{d}{dt}g_k(t)$ are singular at τ_k but integrable suggests that they may be approximated well as a stream of Diracs. Figure 2-6 shows the approximate time derivatives of the scene impulse responses (with singularities removed). Thus, instead of performing spectral estimation using Fourier coefficients $G_k[m]$ of scene impulse response $g_k(t)$, the spectral estimation is done using the Fourier coefficients $i(2\pi m/T)G_k[m]$ of $\frac{d}{dt}g_k(t)$. This is equivalent in the FRI framework to approximating the scene impulse response with 0th order piecewise constant functions.

2.3.2 Recovery Step

The onset times τ_k correspond to times at which the light that travels the shortest total distance is incident on the sensor. Thus from each τ_k we can calculate the shortest path length $\hat{d}_k^{min} = c\tau_k$ for each source-detector pair. Using the previous parameterization of plane $P(\mathbf{n})$ by point \mathbf{n} , let the distance $d_k^{min}(\mathbf{n})$ denote the minimum path length from the origin to $P(\mathbf{n})$ and back to Sensor k at $\mathbf{w}_k = (w_{kx}, w_{ky}, w_{kz})$. From the geometry of the scene, it can be seen that $d_k^{min}(\mathbf{n}) = \|2\mathbf{n} - \mathbf{w}_k\|_2$. This can be used in the following optimization to find the parameters $\hat{\mathbf{n}}$ for which the sum of squared differences between observed total distances and estimated total distances is minimized:

$$\hat{\mathbf{n}} = \underset{\mathbf{n}}{\operatorname{argmin}} \sum_k (d_k^{min}(\mathbf{n}) - \hat{d}_k^{min})^2 \quad (2.5)$$

The resulting plane $P(\hat{\mathbf{n}})$ is the estimate for the plane.

2.3.3 Scene Containing Multiple Planes

Recovering plane orientations for a scene containing multiple intersecting planes can be done in a similar process as previously described in this section. Consider a scene with two planes, P_0 and P_1 , intersecting within the FOV of the camera, as shown earlier in Figure 2-2. The response now has two onset times, τ_{k0} and τ_{k1} ,

corresponding to the times at which the light that travels the shortest total distance to and from each plane is incident on the sensor.

Due to the geometry of the scene, these nearest points will necessarily lie on the visible half planes and not the occluded portions. In addition, the negative contribution from the occluded portions of the plane occur after the initial onsets. Here, the onset times recovered from the initial FRI step need to be assigned to a plane. Assuming that the camera will not be exactly the same distance to both planes, then choose P_0 to be the closer plane. Then the onset times for each plane, τ_{k0} for P_0 and τ_{k1} for P_1 , can be used to recover each plane separately.

2.4 Simulations and Discussion

The simulations demonstrate the performance of the parametric sampling step and explain the parameter choices used for the PSFeaR imaging. The simulations used a 25 cm \times 20 cm device, which is the size of a typical current-generation tablet device. The combined pulse shape and sensor impulse response $h(t)$ was simulated by a Gaussian pulse of width 1 ns. The pulse repetition rate was 50 MHz (signal period $T = 20$ ns) with $N = 201$ samples per repetition period. The signal to noise ratio (SNR) was defined as the maximum ratio over the repetition period between the mean and standard deviation of the approximated Poisson process with rate determined by the scene response, $SNR = \max_n(\sqrt{y_k[n]})$. The SNR was varied for simulations by varying scaling of the scene response, simulating signal dependent noise, and rescaling to the original amplitudes.

2.4.1 Performance of Parametric Sampling Step

The parametric sampling step as discussed in Section 2.3.1 describes the implementation used in the final PSFeaR framework. However, several other factors were taken into consideration for setting parameters such as model order and approximation order and for removing the bias in the estimates before the recovery step.

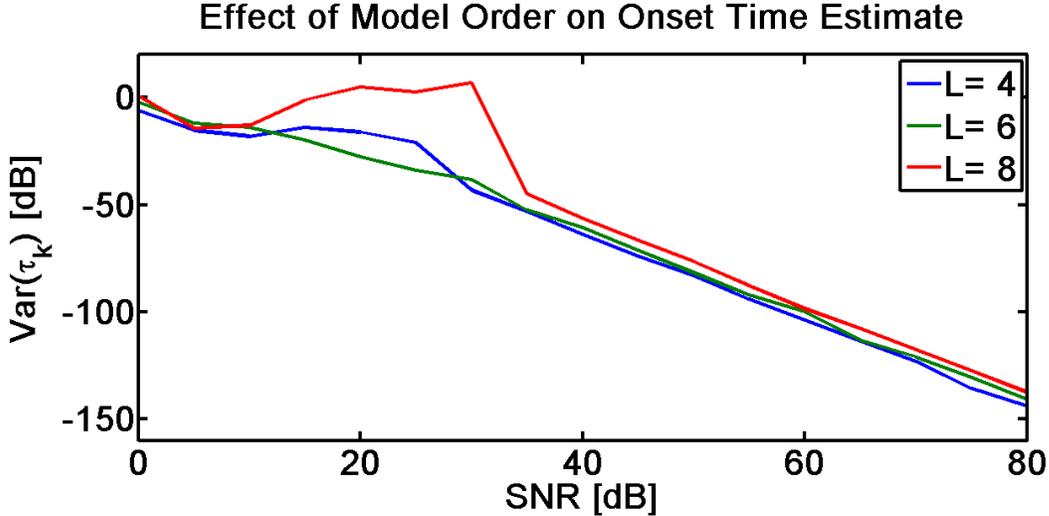


Figure 2-7: Variance of estimates of onset times τ for varying model order with fixed approximation order (0th order box fit)

Model and Approximation Order

Note that for a typical combined pulse shape and sensor impulse response $h(t)$, its Fourier coefficients $H[m]$ for high frequencies have low magnitude relative to coefficients for low frequencies. This deconvolution step in the frequency domain amplifies the effects of high frequency noise. Thus in general there is a tradeoff between using a high enough model order (large enough number L of Fourier coefficients) to fully represent all the Diracs in the signal versus using too high of a model order (too large L) and including the effects of the amplified high frequency noise. Figure 2-7 shows the relative performance of onset time estimation with varying model orders with fixed approximation order over 50 random noisy simulations on the same scene impulse response.

It is possible to also increase the approximation order, or approximate the scene impulse response with higher order piecewise polynomials. While using a 1st order linear spline would intuitively seem to produce a closer reconstruction for the signal, the actual performance in noisy conditions would not reflect the improved modeling as might be expected. Instead, the use of higher order fits increases the effect of high frequency noise, as each additional order introduces another factor of $i(2\pi m/T)$ to the Fourier coefficients used in estimation. Figure 2-8 shows the relative performance

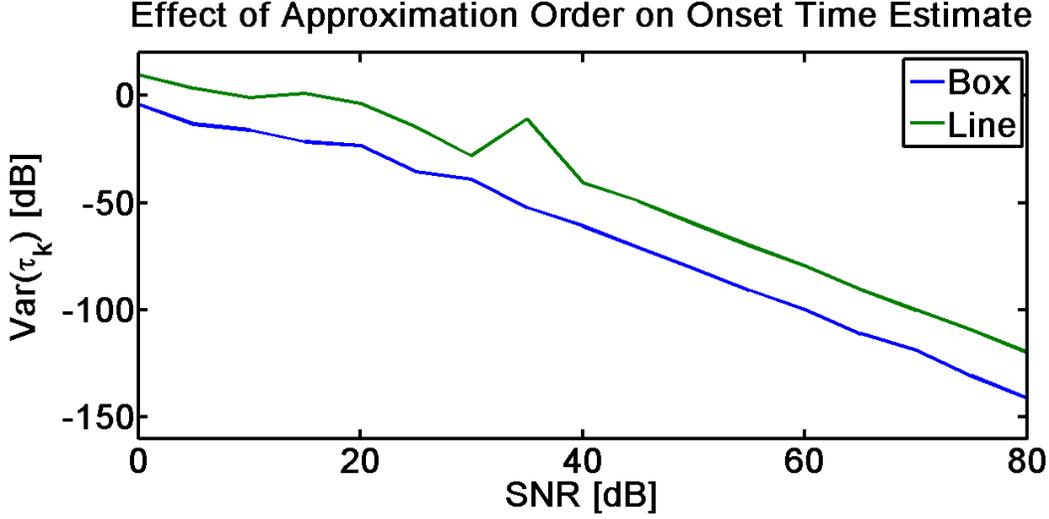


Figure 2-8: Variance of estimates of onset times τ for 0th order (box) and 1st order (linear) approximations for fixed model order $L = 6$

of 0th and 1st order approximations for fixed model order $L = 6$ over 50 random noisy simulations for the same scene impulse response.

Estimation Bias

The derivatives of scene impulse responses are not exactly fit by Diracs and are asymmetric about the singularity. This leads to a slight positive bias when computing locations $\hat{\tau}_k$ of the time locations of the Diracs using the FRI methods. Various factors affecting the bias include true onset time and plane orientation. However, in the typical use case in which the 3D angle θ between the plane normal and the device normal $\langle 0, 0, 1 \rangle$ is less than $\pi/4$, the effect of θ on the bias $b[\hat{\tau}_k; \tau_k, \theta]$ is small and can be marginalized to find an expectation over typical values of θ . Then the bias $b[\hat{\tau}_k; \tau_k] = \hat{\tau}_k - \tau_k$ can be estimated from analysis of scene impulse responses with varying τ_k and corrected. While a closed form expression for the expected bias is difficult to obtain, the correction can be implemented simply via lookup table and interpolation. Figure 2-9 shows the bias of the estimate over typical θ for a range of true τ_k computed via Monte Carlo simulation.

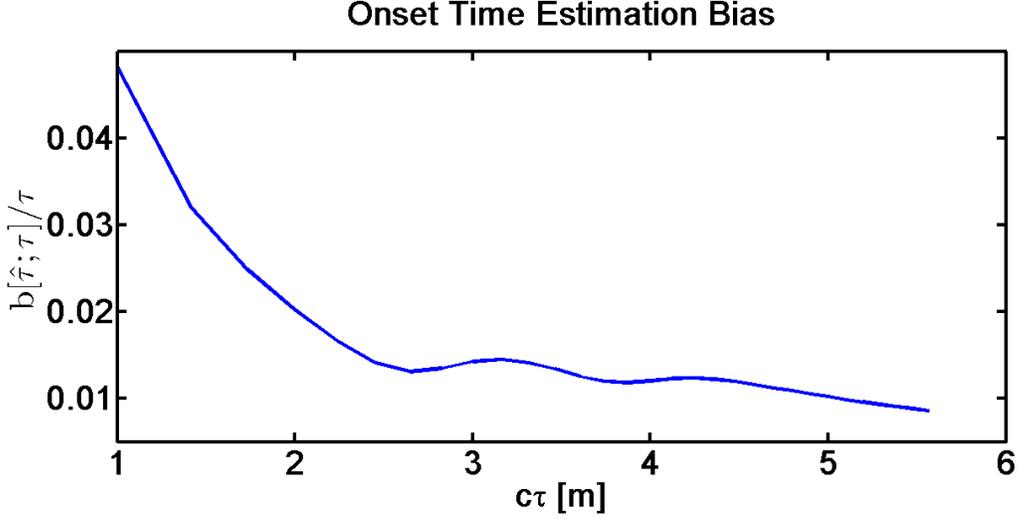


Figure 2-9: Average normalized bias over typical range of scenes

2.4.2 PSFeaR Performance

The performance of the PSFeaR method was compared to the recovery using full forward simulation fitting described previously. The FRI model order was chosen as $L = 4$, and the approximation was made using 0th order piecewise constant fit. To demonstrate the framework, the following two cases were examined:

- a single plane described by the parameterization introduced in the thesis, such that the point in the plane with minimum distance to the origin $\mathbf{n} = (0.4, 0.3, 0.5)$.
- two intersecting planes described by their respective parameterizations $\mathbf{n}_0 = (0, 0.5, 0.5)$ and $\mathbf{n}_1 = (0, -1, 1)$

Figure 2-10 shows the performance of the full forward modeling based recovery method in matching the sampled signal $r_k[n]$ to the expected returns $\hat{y}_k[n]$.

Figures 2-11 and 2-12 show the signal parameter estimation step of the framework for both single and two plane simulations. The estimated signals reconstructed from their parameters are shown with the true signals. Note that although the piecewise-constant fits for $g_k(t)$ are crude due to model mismatch, the important time locations, onset times τ_k , are captured fairly accurately when the time derivative of the scene impulse response $\frac{d}{dt}g_k(t)$ is modeled with Diracs, and the bias can be corrected as described earlier in Section 2.4.1.

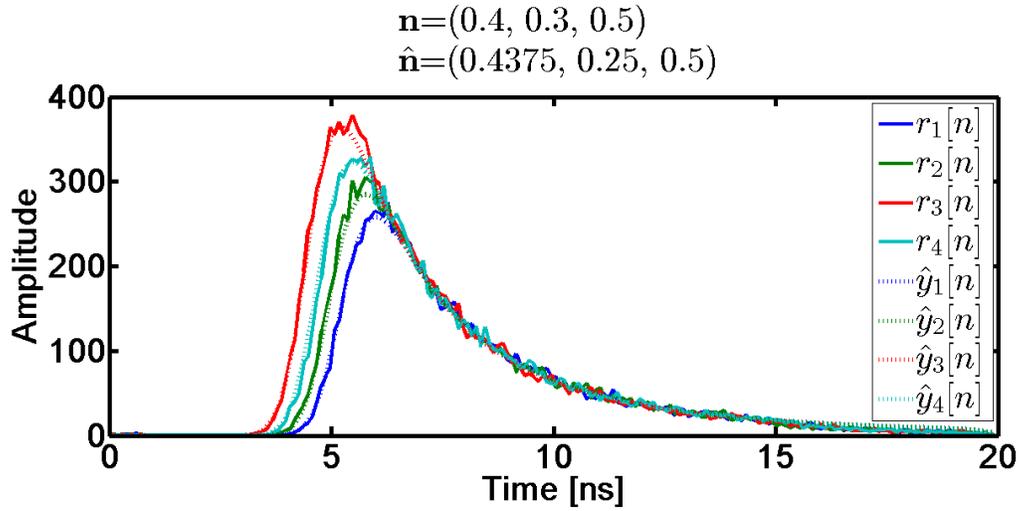


Figure 2-10: Estimates of scene impulse responses using the full forward modeling based recovery.

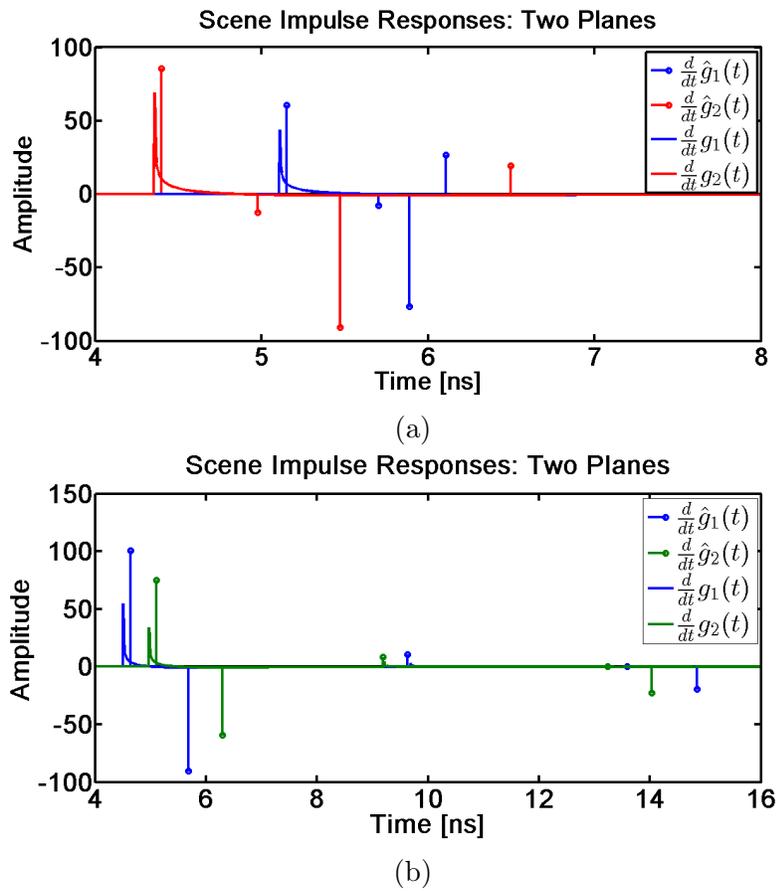
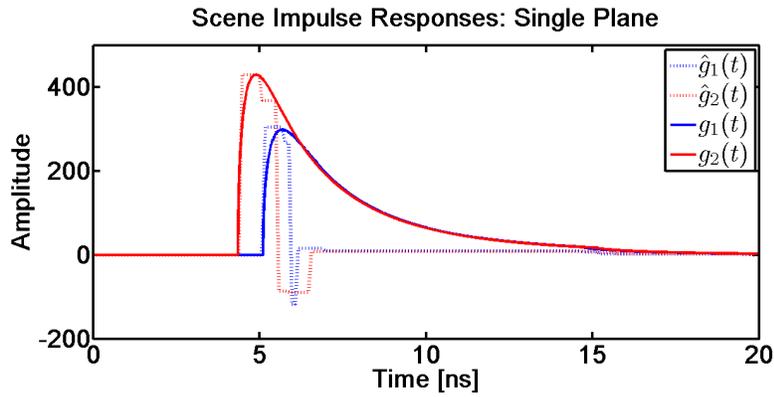
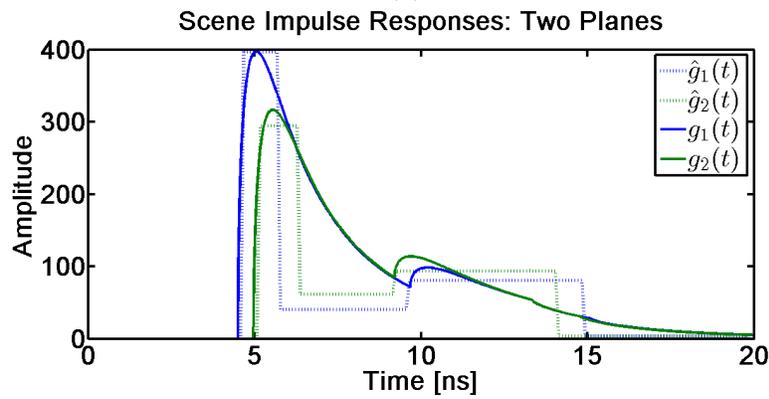


Figure 2-11: Time derivatives and estimated Dirac locations of scene impulse response for simulated scene containing (a) one plane and (b) two planes



(a)



(b)

Figure 2-12: Reconstruction of scene impulse response from estimated Diracs for same simulated scenes as in Figure 2-11 of (a) one plane and (b) two planes.

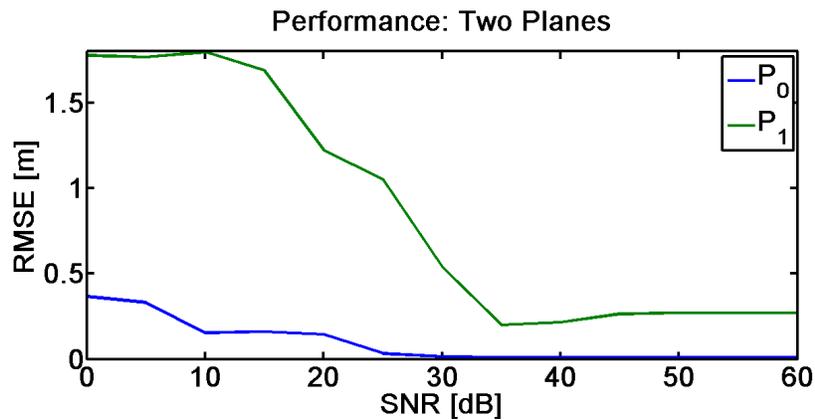


Figure 2-13: Effect of noise on error in estimation of plane parameters for each of two planes. Each noise level was simulated 50 times

Figure 2-13 shows the effects of noise on accuracy for recovering features of each of two planes. For the two plane case, the full forward recovery was infeasible, so the simulations show the ability of the framework to recover multiple planes. It should be noted that the nearer plane is recovered nearly as well as a single plane, but planes with larger distance to the device have weaker returns, so the recovery for these further planes is less accurate. In the simulated case, the second plane was significantly further from the camera than the first plane, and the effect of distance on the accuracy becomes clear.

2.5 Conclusion

Through simulation, it has been demonstrated that the new 3D imaging framework is able to directly estimate features, namely pose and location, from planar scenes by recovering signal parameters without needing to form a full depth map. The two step process is able to recover orientation of multiple planes by fitting scene impulse responses with piecewise constant functions and does so in a tractable computation.

Chapter 3

Simultaneous Phase Unwrapping and Denoising (SPUD)

3.1 Time-of-Flight Imaging

Amplitude modulated cosine wave (AMCW) TOF cameras use a diffuse periodic light source and reflected light focused on a sensor array to measure distance. Because TOF systems use periodic light, the distance measurements they produce can suffer from ambiguities, or phase wrapping. The problem of resolving these ambiguities is a 2D phase unwrapping problem. While there are methods for solving this problem using norm minimization [10], branch cuts [5], and a variety of other techniques, phase unwrapping still remains a significant challenge.

Using multiple modulation frequencies reduces the need for unwrapping by increasing the total unambiguous distance. However, this increases the amount of data collection and acquisition speed required to maintain a fixed frame rate at same noise performance levels. In addition, pixel-wise unwrapping is still not easily or accurately achieved at low SNR.

The Simultaneous Phase Unwrapping and Denoising (SPUD) method is a method based on loopy belief propagation (LBP) addressing both the issues of unwrapping and denoising TOF measurements a unified framework, unlike previous uses of LBP for phase unwrapping [2, 3]. It incorporates an accurate physical model for TOF

measurement and probabilistic scene modeling. It maintains low complexity through the use of generalized approximate message passing (GAMP) and is compatible with current hardware architecture of existing TOF systems. The combined processing that is possible due to the framework allows better performance over postprocessing conventionally formed depth maps.

3.2 Operation of TOF Cameras

The operation of an AMCW TOF cameras make separate measurements for each transverse spatial location (pixel) using an array of sensors. Figure 3-1 shows an abstraction of the signals involved in pixel i when using an AMCW homodyne TOF camera. The periodic source signal $s(t) = 1 + \cos(2\pi ft)$ with modulation frequency $f = 1/T$ and period T is generated from a periodic reference signal $p(t)$ also with period T . The specific design of $p(t)$ will be described shortly. This source $s(t)$ illuminates the entire scene, and the scene return signal at pixel i can be modeled as

$$r_i(t) = a_i \cos(2\pi f(t - \tau_i)) + (a_i + b_i)$$

where a_i is the attenuated amplitude of the reflected sinusoid, $\tau_i = 2z_i/c$ is the time delay due to light travel from camera source to scene point distance z_i and back to camera sensor, c is the speed of light, and b_i is the constant ambient light contribution.

3.2.1 Homodyne Measurement and Conventional Phase Estimation

In homodyne measurement, the return signal $r_i(t)$ is correlated with the phase-shifted reference signal $p(t - \phi)$ and a half-period shifted copy $p(t + T/2 - \phi)$. The significance and choice of the phase shift ϕ will be demonstrated shortly. The signal $p(t)$ is chosen

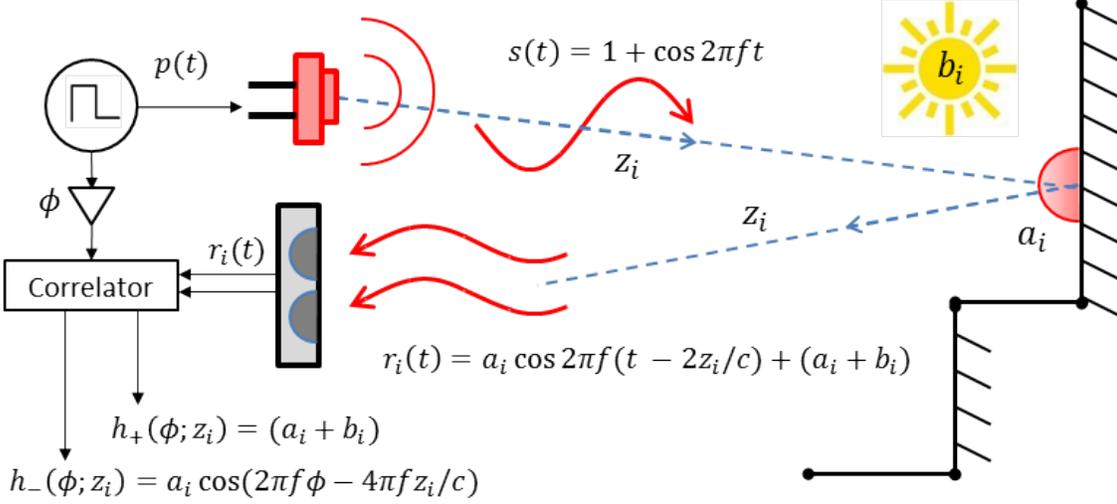


Figure 3-1: Signal processing abstraction of TOF camera

such that the resulting correlation functions have the form,

$$g_-(\phi; z_i) = \frac{1}{T} \int_0^T r_i(t) p(t - \phi) \quad (3.1)$$

$$= a \cos(2\pi f (\phi - 2z_i/c)) + (a + b) \quad (3.2)$$

$$g_+(\phi; z_i) = \frac{1}{T} \int_0^T r_i(t) p(t + T/2 - \phi) \quad (3.3)$$

$$= -a \cos(2\pi f (\phi - 2z_i/c)) + (a + b) \quad (3.4)$$

These two correlations are then added and subtracted (see [4] for details) to yield the final measured outputs as functions of ϕ given,

$$h_+(\phi; z_i) = (g_+(\phi; z_i) + g_-(\phi; z_i))/2 = (a_i + b_i) \quad (3.5)$$

$$h_-(\phi; z_i) = (g_+(\phi; z_i) - g_-(\phi; z_i))/2 = a_i \cos(2\pi f \phi - 4\pi f z_i/c) \quad (3.6)$$

The importance of phase shift ϕ of the reference signal $p(t - \phi)$ and $p(t + T/2 - \phi)$ becomes clear. The correlation functions $h_+(\phi; z_i)$ can be effectively sampled by varying ϕ . Estimating the amplitude and frequency of the sinusoidal signal $h_-(\phi; z_i)$ requires at least two samples in ϕ per period; however, typical TOF cameras collect

4 equally spaced samples $y_i^{(n)} = h_{i+}(\phi_n; z_i)$ in each period, $\phi_n = n/4f$ for $n = 0, 1, 2, 3$. The remainder of this thesis will assume a 4-tuple of measurements $\mathbf{y}_i = (y_i^{(0)}, y_i^{(1)}, y_i^{(2)}, y_i^{(3)})$, is obtained at each pixel i at depth z_i using modulation frequency f . In addition, the camera is assumed to similarly produce samples $h_+^{(n)}$ of $h_+(\phi_n; z_i)$. From these measurements, conventional processing in current TOF cameras gives pointwise estimates for sinusoid amplitude, background contribution, and wrapped distance:

$$\tilde{a}_i = \sqrt{\frac{(y_i^{(0)})^2 + (y_i^{(1)})^2 + (y_i^{(2)})^2 + (y_i^{(3)})^2}{2}}, \quad (3.7)$$

$$\tilde{b}_i = \frac{1}{4}(h_+^{(0)} + h_+^{(1)} + h_+^{(2)} + h_+^{(3)}) - \tilde{a}_i, \quad (3.8)$$

$$\tilde{z}_i = \frac{c}{4\pi f} \tan^{-1} \left(\frac{y_i^{(1)} - y_i^{(3)}}{y_i^{(0)} - y_i^{(2)}} \right). \quad (3.9)$$

These estimates are accurate in the absence of measurement noise.

3.2.2 Statistical Measurement

As discussed in Chapter 1, the process of converting photons to electric charge in photodetector sensors generates shot noise [8], which affects TOF camera measurements. As described previously, the charge generation can be modeled as a time inhomogeneous Poisson process with rate $\lambda_i(t) = \eta r_i(t)$ at each sensor pixel, where η is the quantum efficiency of the sensor.

To combat the effects of this shot noise, the correlation function is averaged over N periods. For large N , according to the central limit theorem, the output is approximately normally distributed,

$$y_i^{(n)} | z_i \sim \mathcal{N} \left(\eta g \left(\frac{n}{4f}; z_i \right), \frac{1}{2} \eta (a_i + b_i) / N \right) \quad \text{for } n = 0, 1, 2, 3.$$

Given z_i , each measurement of the correlation function is conditionally independent. Thus the joint distribution or the 4-tuple of measurements \mathbf{y}_i given true distance z_i

at pixel i is given

$$\begin{aligned} p(\mathbf{y}_i | z_i) &\propto \prod_{n=0}^3 \exp \left\{ - \frac{(\eta a_i \cos(\frac{1}{2}n\pi - 4\pi f c^{-1} z_i) - y_i^{(n)})^2}{2\sigma_i^2} \right\} \\ &\propto C \exp \left\{ \sum_{n=0}^3 \frac{\eta a_i y_i^{(n)} \cos(\frac{1}{2}n\pi - 4\pi f c^{-1} z_i)}{\sigma_i^2} \right\}, \end{aligned}$$

where $\sigma_i^2 = \frac{1}{2}\eta(a_i + b_i)/N$. Then the likelihood on observations \mathbf{y}_i given true distance z_i takes the form

$$p(\mathbf{y}_i | z_i) = C \exp \left\{ \frac{\eta a_i A_i}{\sigma_i^2} \cos(4\pi f c^{-1}(z_i - \tilde{z}_i)) \right\}, \quad (3.10)$$

where $A_i = \sqrt{(y_i^{(0)} - y_i^{(2)})^2 + (y_i^{(1)} - y_i^{(3)})^2}$ and \tilde{z}_i is defined as in (3.9), which is available from measurements taken by current TOF cameras. Since a_i and b_i are not available directly to form our likelihood, we instead use the values \tilde{a}_i and \tilde{b}_i estimated as in (3.7) – (3.8), which are also readily available in current TOF cameras.

3.2.3 Phase Unwrapping

Because TOF cameras use periodic illumination signals, measurements beyond a certain range are ambiguous, and many different true depth values may produce the same measurement. In particular, $z_i = \tilde{z}_i + nc/(2f)$ for any $n \in \mathbb{Z}$. Decreasing the modulation frequency decreases the maximum unambiguous distance $D = c/2f$. However, it is well-known that with noisy measurements, the error in estimating distance increases as well. This fact can be derived from the earlier analysis.

With two different modulation frequencies for illumination f_0 and f_1 ,

$$z_i = \tilde{z}_{i0} + n_0 \frac{c}{2f_0} = \tilde{z}_{i1} + n_1 \frac{c}{2f_1} \quad \text{for } n_0, n_1 \in \mathbb{Z} \quad (3.11)$$

is the system of equations determining the ambiguity of measurements, where \tilde{z}_{ij} is the measured distance at frequency f_j . It can be seen that the maximum unambiguous distance is extended to $D = c/2 \gcd(f_0, f_1)$ when f_0 and f_1 are chosen to have some

common factors. In the absence of noise, the true distance within the can be computed accurately from 3.11. However, this algebraic system is not robust, and small errors in measurement can result in large inaccuracies in estimated unwrapped distance. Thus in a noisy setting, the unwrapping problem becomes nontrivial.

Consider the joint likelihood of measuring two 4-tuples at each of two modulation frequencies, $p(\mathbf{y}_{i0}, \mathbf{y}_{i1}|z_i) = p(\mathbf{y}_{i0}|z_i)p(\mathbf{y}_{i1}|z_i)$. The value of z_i that maximizes this likelihood can be used as a pointwise estimate for the tru distance of the scene point. However, this likelihood is highly nonconvex, and finding such a maximum likelihood estimate (MLE) essentially requires a global grid search within the extended unambiguous range.

3.3 Unwrapping and Denoising TOF Measurements

Unwrapping using two frequencies was a nontrivial problem in noisy settings. In this section, a graphical modeling based framework for solving the unwrapping problem with noise will be introduced.

3.3.1 Transform Domain Signal Modeling

As discussed in Chapter 1, images of natural scenes often have sparse representations in discrete wavelet bases, with wavelet coefficients following approximately Laplacian distributions,

$$p(x_k) = \frac{1}{2q} e^{-|x_k/q|} \quad (3.12)$$

for some parameter q that varies between depth maps based on scene structure. This knowledge of sparsity can be used to denoise images. For example, one method for denoising is to perform soft thresholding on the wavelet coefficients of an image, which is equivalent to performing the maximum a posteriori probability estimate with a Laplacian prior on the coefficients assuming a Gaussian noise model.

Pairing this notion of wavelet sparsity with the previous description of the TOF measurement and noise processes, the forward acquisition can be represented as a

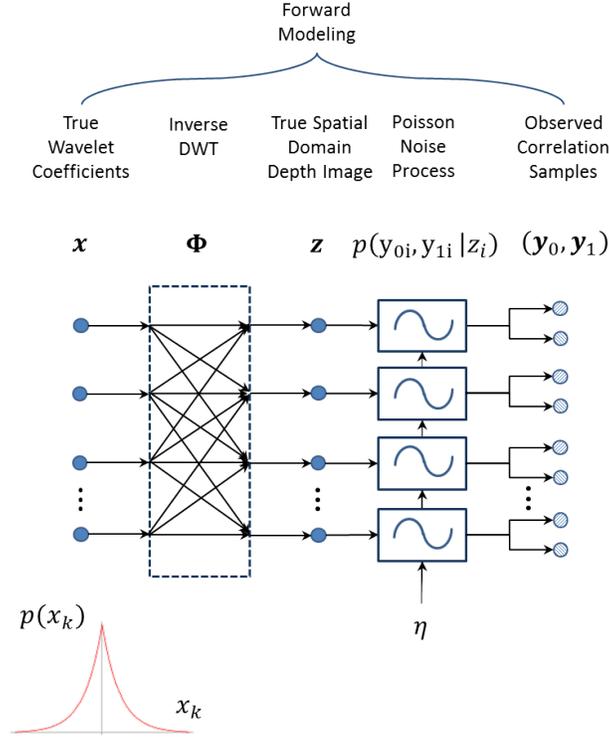


Figure 3-2: Forward acquisition model for the TOF camera measurements at two modulation frequencies.

graphical model as in Figure 3-2.

3.3.2 GAMP-Based Unwrapping and Denoising

Fig. 3-2 shows a graphical model for the acquisition process. This is amenable to GAMP since wavelet coefficients \mathbf{x} can be modeled as independent and the measurement process is separable on the spatial-domain quantities $\mathbf{z} = \Phi \mathbf{x}$. Like approximate message passing [1] and earlier techniques, GAMP uses approximations for certain messages in loopy BP; unlike earlier techniques it allows the non-Gaussianity of noise and nonlinearities discussed in Section 3.2.

To incorporate our model into the GAMP framework, the prior distributions $p(x_k)$ and probabilistic measurement channels $p(\mathbf{y}_{0i}, \mathbf{y}_{1i} | z_i)$ as well as the linear transform matrix Φ need to be specified. The distribution $p(x_k)$ given in (3.12) enforces sparsity of the wavelet coefficients of the estimate as discussed in Chapter 1. The linear mixing matrix Φ performs the inverse discrete wavelet transform on \mathbf{x} to obtain

the unwrapped image $\mathbf{z} = \Phi \mathbf{x}$. The measurement channel probability distribution $p(\mathbf{y}_{0i}, \mathbf{y}_{1i} | z_i)$ incorporates the observed data into the estimate. The algorithm is summarized below in Alg. 1, and one iteration is illustrated in Fig. 3-3.

Algorithm 1 (High-level pseudocode for GAMP algorithm)

initialize: $\mathbf{x}(0) = \mathbf{0}, \mathbf{z}(-1) = \mathbf{0}$
for $t = 0 \rightarrow N - 1$ **do**

1. *Output linear step:* Compute
 - $\hat{\boldsymbol{\mu}}_z(t)$ from $\Phi, \boldsymbol{\mu}_x(t)$, and $\boldsymbol{\mu}_z(t - 1)$;
 - $\hat{\boldsymbol{\sigma}}_z(t)$ from Φ and $\boldsymbol{\sigma}_x(t)$
2. *Output nonlinear step:* For each i , estimate
 - $(\mu_{z_i}(t), \sigma_{z_i}(t))$ from
 - $(\hat{\mu}_{z_i}(t), \hat{\sigma}_{z_i}(t))$ and $p(\mathbf{y}_{0i}, \mathbf{y}_{1i} | z_i)$
3. *Input linear step:* Compute
 - $\hat{\boldsymbol{\mu}}_x(t)$ from Φ^T and $\boldsymbol{\mu}_z(t)$;
 - $\hat{\boldsymbol{\sigma}}_x(t)$ from Φ^T and $\boldsymbol{\sigma}_z(t)$
4. *Input nonlinear step:* For each k , estimate
 - $(\mu_{x_k}(t + 1), \sigma_{x_k}(t + 1))$ from
 - $(\hat{\mu}_{x_k}(t), \hat{\sigma}_{x_k}(t))$ and $p(x_k)$

end for

The “sum-product” and “max-sum” versions of GAMP described in [12] both follow this outlined algorithm, employing different estimator forms in the output and input nonlinear steps. The output and input nonlinear steps for the “sum-product” version as well as computationally-efficient approximations for these steps will be described in more detail in Sections 3.3.3 and 3.3.4.

As outlined, the GAMP framework gives as output an estimate of wavelet coefficients \mathbf{x} , which through the linear transform Φ is used to find the unwrapped depth image estimate $\mathbf{z} = \Phi \mathbf{x}$. This outlined method will be called “SPUD1.” Alternatively, an additional half iteration can be performed to estimate \mathbf{z} from the output nonlinear step with wavelet coefficients \mathbf{x} as input. This corresponds to further performing steps (1) and (2) from Algorithm 1 beyond SPUD1. This modified method

will be known as “SPUD2.”

It is observed that the two estimates from using SPUD1 and SPUD2 do not necessarily converge to a single point, but instead can reach a fixed cycle in which each method converges to its own distinct equilibrium point. The SPUD1 estimate is more strongly affected by the sparse prior, while the SPUD2 estimate is more influenced by the observed data. Thus with the same sparsity model and filter length, when more denoising is required in low SNR settings, SPUD1 outperforms SPUD2, but when the observations are more reliable in higher SNR settings, SPUD2 tends to produce more exact estimates of the scene.

3.3.3 Approximating the Output Nonlinear Step

The output nonlinear step in Algorithm 1 estimates an expectation and variance for z_i based on the sample vectors $(\mathbf{y}_{0i}, \mathbf{y}_{1i})$, the likelihood function $p(\mathbf{y}_{0i}, \mathbf{y}_{1i}|z_i)$, and the previous computed values for the expectation and variance of \hat{z}_i .

The calculation of the estimates $\mu_{z_i}(t)$ and $\sigma_{z_i}(t)$ are

$$\begin{aligned}\mu_{z_i}(t) &= \int_{-\infty}^{\infty} z f(z, \hat{\mu}_{z_i}(t), \hat{\sigma}_{z_i}(t)) dz \\ \sigma_{z_i}(t) &= \int_{-\infty}^{\infty} (z - \mu_{z_i}(t))^2 f(z, \hat{\mu}_{z_i}(t), \hat{\sigma}_{z_i}(t)) dz\end{aligned}$$

where

$$f(z, \hat{\mu}_{z_i}(t), \hat{\sigma}_{z_i}(t)) = p(\mathbf{y}_{0i}, \mathbf{y}_{1i}|z) \mathcal{N}(z; \hat{\mu}_{z_i}(t), \hat{\sigma}_{z_i}(t)) \quad (3.13)$$

(see [11] for further details).

However, the exact integral does not have a closed-form expression and is cumbersome to numerically integrate. Instead, an approximation to the likelihood function can be made to simplify computation. The exponential of a cosine from (3.10) may be closely approximated by the wrapped normal distribution [15]. For $j = 0, 1$ and

each i , the approximated likelihood function is

$$\tilde{p}(\mathbf{y}_{ji}|z_i) \propto \sum_{n=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_{ji}} \exp \left\{ -\frac{(z_i - \tilde{z}_{ji} - nD_j)^2}{2\sigma_{ji}^2} \right\}$$

where

$$\sigma_{ji}^2 = \frac{2c^2}{(4\pi f_j)^2} \ln \left(\frac{I_0(\eta a_i A_{ji}/\sigma_i^2)}{I_1(\eta a_i A_{ji}/\sigma_i^2)} \right)$$

and $I_n(k)$ is the modified Bessel function of the first kind. The likelihoods $p(\mathbf{y}_{0i}|z_i)$ and $p(\mathbf{y}_{1i}|z_i)$ can then be further approximated without incurring much additional error within the unambiguous range by truncation to sums of L and M Gaussians respectively. The resulting approximation for $p(\mathbf{y}_{0i}, \mathbf{y}_{1i}|z_i) = p(\mathbf{y}_{0i}|z_i)p(\mathbf{y}_{1i}|z_i)$ can be expressed as a weighted sum of $K = LM$ Gaussians,

$$\tilde{p}(\mathbf{y}_{0i}, \mathbf{y}_{1i}|z_i) \approx \sum_{k=1}^K w_k \mathcal{N}(z_i; \tilde{s}_k, \sigma_k^2)$$

Now $f(z, \hat{\mu}_{z_i}(t), \hat{\sigma}_{z_i}(t))$ from (3.13) can be expressed as a weighted sum of K Gaussians, and the computation for estimates $\mu_{z_i}(t)$ and $\sigma_{z_i}(t)$ can be simply characterized by the values of w_k , \tilde{s}_k , σ_k^2 , $\hat{\mu}_{z_i}(t)$, and $\hat{\sigma}_{z_i}(t)$.

To evaluate the performance of the approximation for the exponential of a cosine, the error between the function and its approximation as a function of k , the coefficient for the cosine in the likelihood function for a single frequency, is calculated over the extended unambiguous range as

$$\begin{aligned} e_j(k) &= \left\| \exp \left(k \cos \frac{2\pi z}{D_j} \right) - \tilde{p}_j(k, z) \right\|_2 \\ &= \left(\int_{-D/2}^{D/2} \left| \exp \left(k \cos \frac{2\pi z}{D_j} \right) - \tilde{p}_j(k, z) \right|^2 dz \right)^{1/2} \end{aligned}$$

where $\tilde{p}_j(k, z)$ is the approximation to the single frequency likelihood. The error for the two frequency likelihood over the extended unambiguous range can be calculated

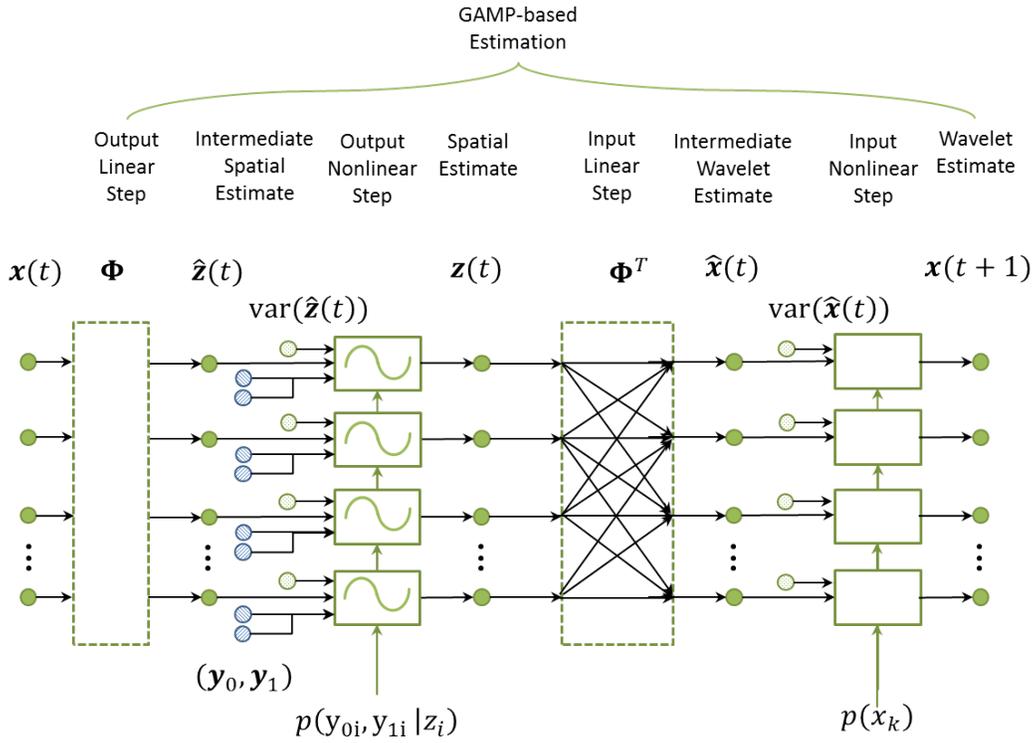


Figure 3-3: One iteration of our GAMP algorithm for unwrapping and denoising. This diagram shows the updates of the estimates for $\hat{z}(t)$ and $\hat{x}(t)$, represented by dark green nodes. The means of the estimates $\hat{\mu}_z(t)$ and $\hat{\mu}_x(t)$ and the variances $\hat{\sigma}_z(t)$ and $\hat{\sigma}_x(t)$ are updated concurrently at each step

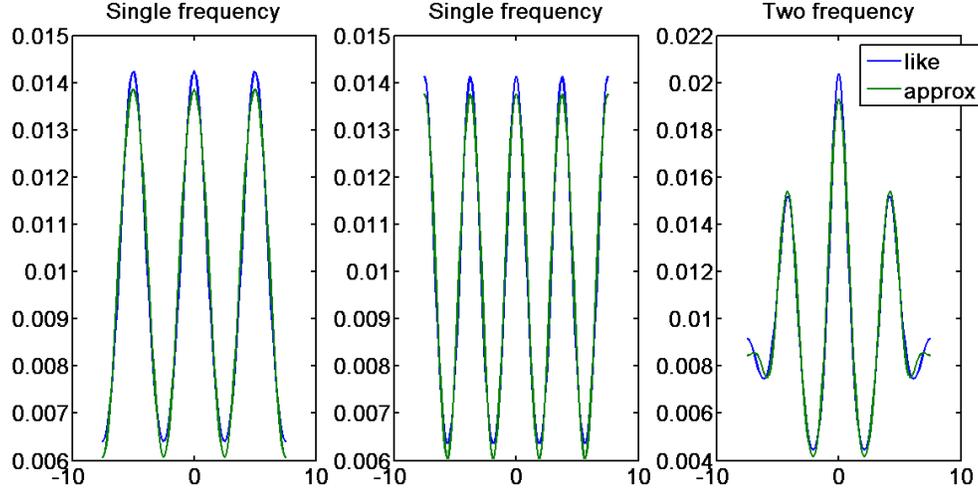


Figure 3-4: Differences between sum of Gaussian approximation and exponential of cosine functions for low frequency, high frequency, and two frequencies at $k = 0.4$

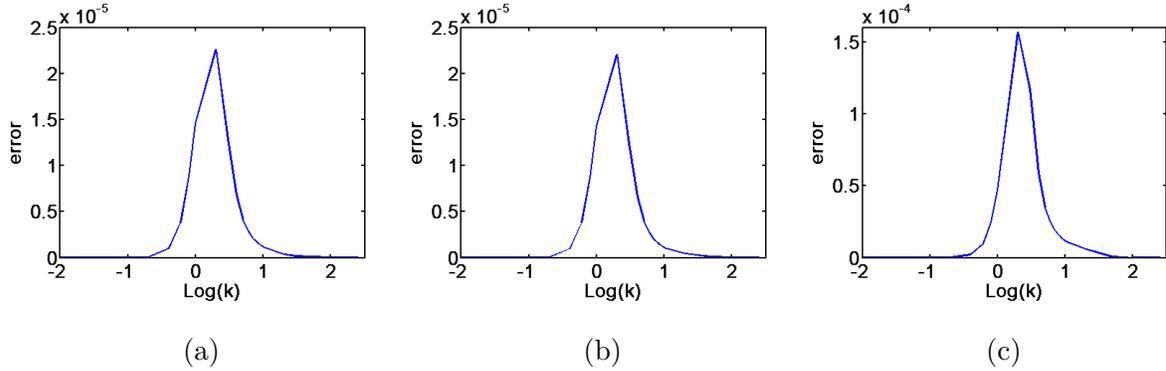


Figure 3-5: Errors of sum of Gaussian approximation for exponential of cosine likelihood functions for (3-5a) low frequency, (3-5b) high frequency, and (3-5c) two frequencies over a range of k values

similarly,

$$e(k) = \left\| \exp\left(k \cos \frac{2\pi z}{D_0} + k \cos \frac{2\pi z}{D_1}\right) - \tilde{p}_0(k, z) \tilde{p}_1(k, z) \right\|_2$$

Figure 3-5 shows the error in approximating each single frequency likelihood separately and the error for approximating the combined likelihood.

3.3.4 Approximating the Input Nonlinear Step

The input nonlinear step in Algorithm 1 estimates an expectation and variance for x_k based on the prior distribution $p(x_k)$ and the previous computed values for expectation and variance for \hat{x}_k . This calculation also requires computing similar integrals for $\mu_{x_k}(t)$ and $\sigma_{x_k}(t)$,

$$\begin{aligned}\mu_{x_k}(t) &= \int_{-\infty}^{\infty} xg(x, \hat{\mu}_{x_k}(t), \hat{\sigma}_{x_k}(t)) dx \\ \sigma_{x_k}(t) &= \int_{-\infty}^{\infty} (x - \mu_{x_k}(t))^2 g(x, \hat{\mu}_{x_k}(t), \hat{\sigma}_{x_k}(t)) dx\end{aligned}$$

where

$$g(x, \hat{\mu}_{x_k}(t), \hat{\sigma}_{x_k}(t)) = p(x_k)\mathcal{N}(x; \hat{\mu}_{x_k}(t), \hat{\sigma}_{x_k}(t)) \quad (3.14)$$

While these integrals can be expressed in terms of exponentials and the error function $\text{erf}(\cdot)$, the expressions are not numerically well-behaved. Instead, an approximation can be made here as well, replacing the expectation with a more well-behaved maximum a posteriori estimate, and a corresponding variance update

$$\begin{aligned}\mu_{x_k}(t) &= \underset{x_k}{\text{argmax}} g(x, \hat{\mu}_{x_k}(t), \hat{\sigma}_{x_k}(t)) \\ \sigma_{x_k}(t) &= \hat{\sigma}_{x_k}(t)\end{aligned}$$

3.3.5 Computation

The use of the Gaussian mixture to approximate the true likelihood leads to a simple form for the expectation and variance computations in the output nonlinear step—much lower than numerical integration using the true likelihood. The linear mixing can be implemented using a 2D fast discrete wavelet transform. Furthermore, the structure of the algorithm allows the processing to be parallelized and optimized for computation in real time, implemented either in hardware or on a GPU, for example. Thus, the full integrated processing can be performed quickly and accurately.

3.4 Simulations and Discussion

The integrated SPUD methods were compared against separate pointwise maximum likelihood estimation for unwrapping followed by wavelet thresholding and other conventional techniques for denoising. The modulation frequencies chosen were 30 MHz and 40 MHz, typical operating frequencies for TOF cameras, yielding a maximum unambiguous range of $c/(2 \cdot 10 \text{ MHz}) \approx 15 \text{ m}$ from the gcd modulation frequency of 10 MHz. The simulated scenes were within the range of 0.5–12 m, causing wrapping for either modulation taken separately.

The simulated scenes were taken from the right side camera of the Tsukuba stereo pair dataset under different illumination conditions [7], [9]. Ground truth depth maps simulated z , the flashlight illumination simulated active illumination and spatially varying reflectivity in amplitude a , and fluorescent light illumination simulated ambient background contribution b . Defining the SNR to be the ratio between the average amplitude of the sinusoid and the standard deviation of noise in the samples, $\text{SNR} = 10 \log_{10}(2a^2N/(a+b)) \text{ dB}$, and holding constant the integration periods $N = 100$, we vary the SNR by varying the average level of b .

The 2D separable Daubechies length-4 discrete wavelet transform was used in both the baseline methods and the SPUD methods. SPUD was run for 20 iterations with step size 0.5, and the MLE was found using a discretized grid search. Wiener filtering was also performed using the estimated spectrum and noise variance of the MLE depth map. Median filtering was performed for comparison as well.

Fig. 3-6 shows the “Tsukuba 450” simulated scene along with the wrapped images produced by using (3.9) separately for each modulation frequency, with $\text{SNR} = 10 \text{ dB}$. Pointwise ML estimation is difficult (requiring a grid search) and does not perform well, as shown in Fig. 3-7(a,b). Post-processing the MLE by wavelet thresholding (using MatLab’s default thresholding function) provides significant improvement, as shown in Fig. 3-7(c,d). Wiener filtering provides similar MSE performance, as shown in Fig. 3-7(e,f). While wavelet thresholding oversmooths most of the image, the adaptive Wiener filter oversmooths in some blocks while leaving some of the noisier

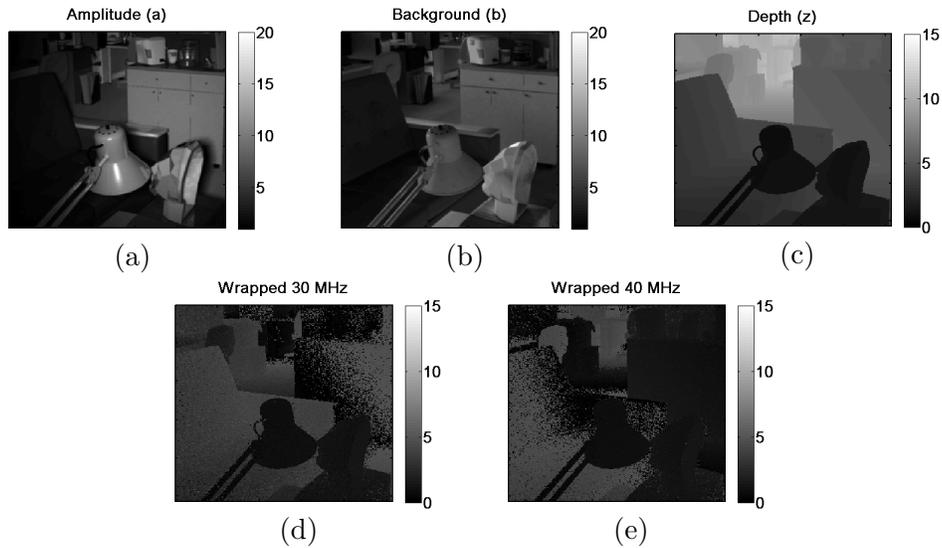


Figure 3-6: Ground truth and estimates for “Tsukuba 450” scene computed from (3.9) using single modulation frequencies at SNR = 10 dB.

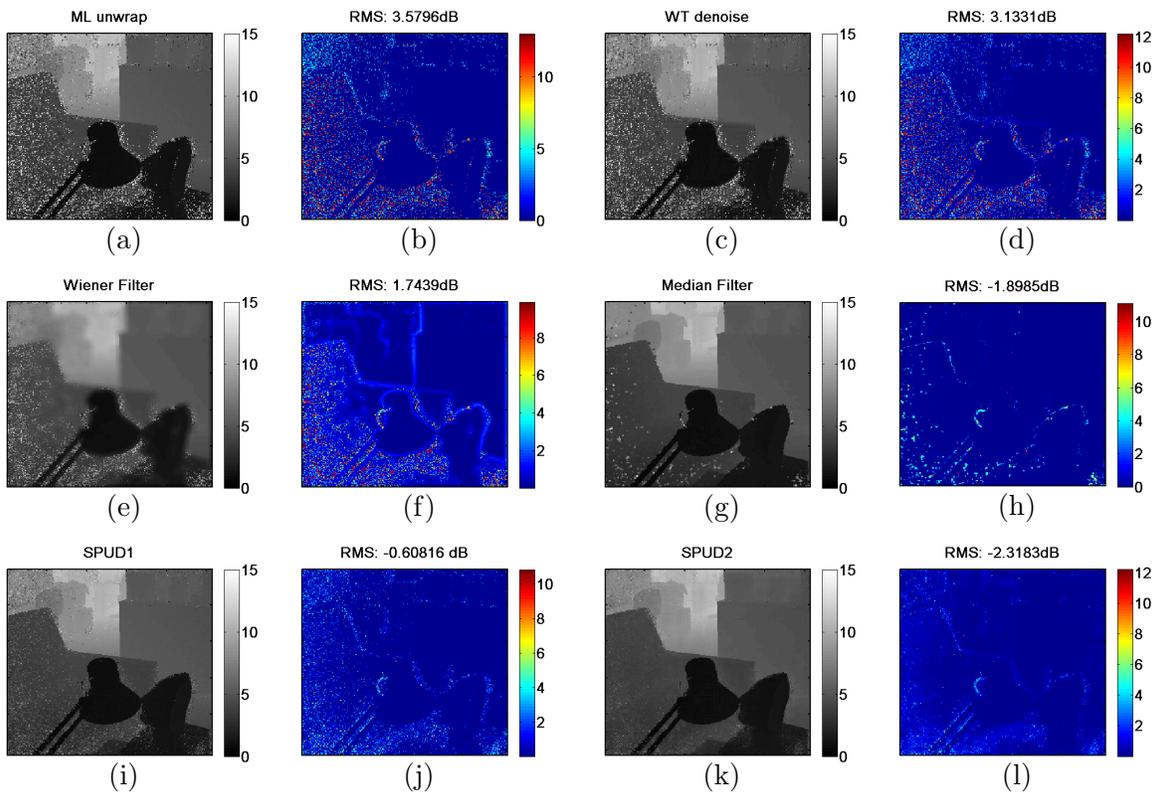


Figure 3-7: Reconstructed depth maps and their root mean squared errors. The proposed GAMP method provides a 0.5 dB improvement relative to the median filtered MLE and more than 5 dB relative to the pointwise MLE. All images produced at SNR = 10 dB.

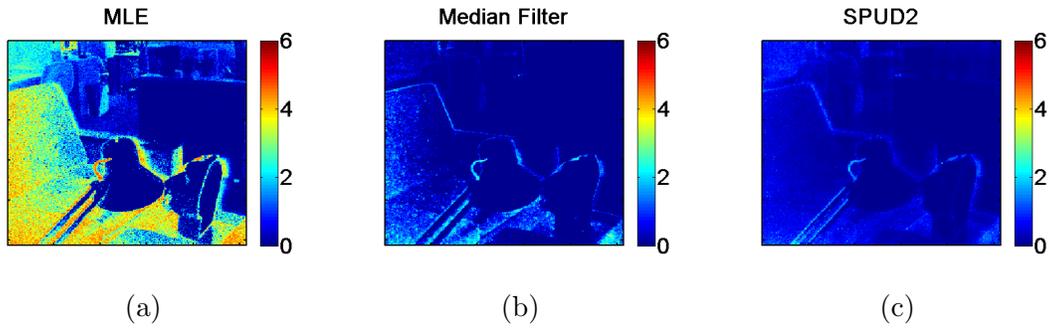


Figure 3-8: Pixelwise standard deviations over 50 simulations at 10 dB SNR

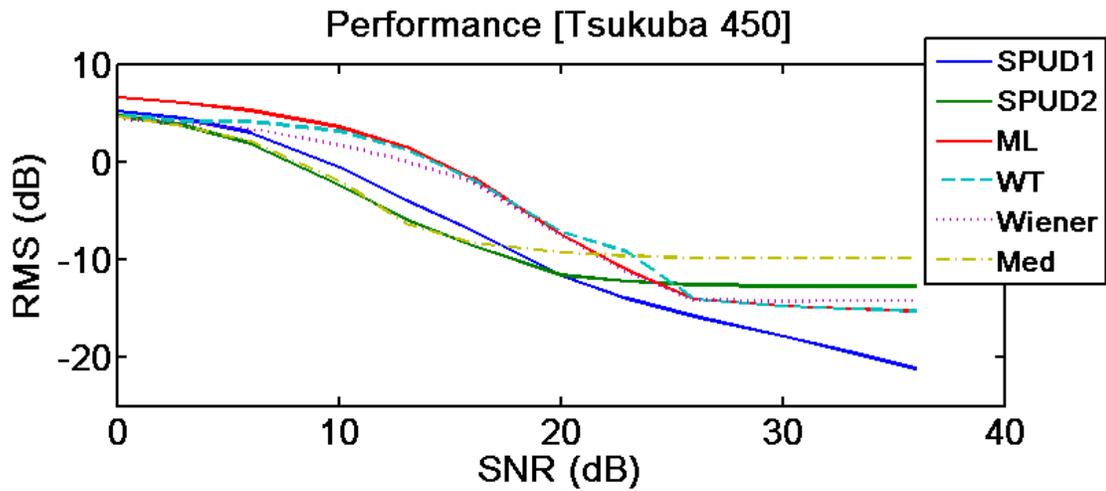


Figure 3-9: Root mean squared error comparison for several methods across an SNR range for “Tsukuba 450” image.

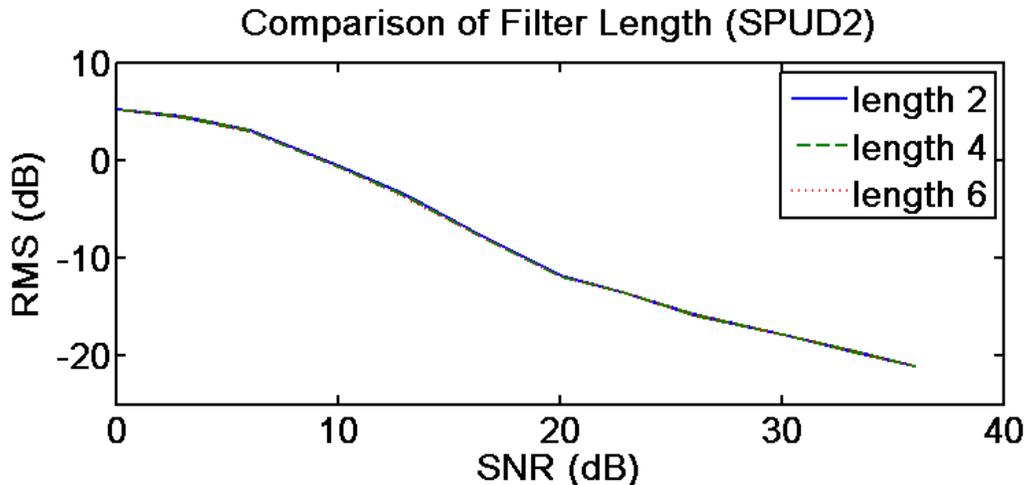


Figure 3-10: Comparison of filter lengths on “Tsukuba 450” image across an SNR range.

patches untouched. The median filter performs well at very low SNR but not across the entire SNR range and also produces large blocks of incorrect estimates. The proposed GAMP-based method does much better than the post-processing, both visually and in RMS, as shown in Fig. 3-7(g,h).

Fig. 3-8 shows the per-pixel standard deviations achieved over 50 Monte Carlo simulations at 10 dB SNR using the MLE, median filtering on the MLE, and SPUD2. The standard deviations vary with the illumination intensities. Median filtering provides a reduction in standard deviation over the MLE estimate, but SPUD2 yields lower standard deviation across the depth image than median filtering.

Fig. 3-9 provides a comparison over a range of low and moderate SNRs. At extremely low SNRs, all methods fail to produce a useful image; the proposed method allows successful operation at moderate SNRs, for which we can increase robustness to ambient light, lower the illumination power requirement, or increase the frame rate.

Fig. 3-10 shows the effect of filter length on the GAMP-based reconstruction performance. The length-4 Daubechies filter provides reasonable improvement over length-2 filter at low and moderate SNRs, but further increasing the length does not result in much more significant reduction in MSE. All simulation code and simulation data are available at <http://rleweb.mit.edu/stir/spud/>.

Chapter 4

Conclusion

4.1 PSFeaR

The PSFeaR framework was demonstrated to be able to practically and directly estimate features of planar scenes by recovering signal parameters. The two step process finds plane locations and orientations for scenes composed of single and multiple planes by approximating the time derivatives of scene impulse responses by Diracs. This architecture processing greatly reduces the complexity of 3D acquisition by directly inferring features rather than forming a full high resolution depth map.

The signal parameter recovery in Step 1 of the framework is somewhat susceptible to noise for more distant or less reflective planar scenes due to the model mismatch. In addition, the PSFeaR framework estimates scene features assuming that recovered signal parameters vary according to a normal distribution, which is not necessarily the best fit based on the parameter recovery method. Future work includes increasing the parametric deconvolution performance in Step 1 and incorporating the distribution of recovered parameters when estimating scene features in Step 2. Additionally, the effects of a spatially varying reflectivity function as well as varying combined pulse shape and sensor impulse responses on the scene impulse response and its recovery could be studied in more detail to better inform the modeling of the recovered parameter distribution.

4.2 SPUD

SPUD is a new method for integrated processing of data in time-of-flight cameras to perform unwrapping and denoising jointly. The result is greatly improved performance over post-processing of a noisy unwrapped image, particularly at low SNRs. Because the scene is within the extended unambiguous range, the success of the current method is due largely to the detailed acquisition modeling rather than the use of a signal prior.

Future work could extend the integrated approach to image beyond the unambiguous range, possibly even using a single modulation frequency, for particular classes of scenes. Parameters of a separable image prior could be automatically estimated by incorporating the Expectation-Maximization algorithm [18]. Greater structure in the signal prior could be incorporated using hybrid GAMP [12].

Bibliography

- [1] D. L. Donoho, A. Maleki, and A. Montanari. Message-passing algorithms for compressed sensing. *Proc. Nat. Acad. Sci.*, 106(45):18914–18919, November 2009.
- [2] D. Droeschel, D. Holz, and S. Behnke. Multi-frequency phase unwrapping for time-of-flight cameras. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots & Syst.*, pages 1463–1469, Taipei, Taiwan, October 2010.
- [3] B. J. Frey, R. Koetter, and N. Petrovic. Very loopy belief propagation for unwrapping phase images. In T. Diettrich (Author), S. Becker, and Z. Ghahramani, editors, *Proc. Neural Information Process. Syst.*, Vancouver, Canada, December 2001.
- [4] S. Burak Gokturk, Hakan Yalcin, and Cyrus Bamji. A time-of-flight depth sensor — system description, issues and solutions. In *Proc. Conf. Comput. Vis. Pattern Recog. Workshop*, page 35, 2004.
- [5] R. M. Goldstein, H. A. Zebker, and C. L. Werner. Satellite radar interferometry: Two-dimensional phase unwrapping. *Radio Sci.*, 23(4):713–720, January 1988.
- [6] A. Kirmani, H. Jeelani, V. Montazerhodjat, and V. K. Goyal. Diffuse imaging: Creating optical images with unfocused time-resolved illumination and sensing. *IEEE Signal Process. Lett.*, 19(1):31–34, January 2012.
- [7] Sarah Martull, Martin Peris, and Kazuhiro Fukui. Realistic cg stereo image dataset with ground truth disparity maps. *ICPR workshop TrakMark2012*, 111(430):117–118, 2012.
- [8] F. Mufti and R. Mahony. Statistical analysis of measurement processes for time-of-flight cameras. In *Proc. SPIE*, volume 7447, page 74470I, 2009.
- [9] Martin Peris, Sara Martull, Atsuto Maki, Yasuhiro Ohkawa, and Kazuhiro Fukui. Towards a simulation driven stereo vision system. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 1038–1042. IEEE, 2012.
- [10] M. D. Pritt and J. S. Shipman. Least-squares two-dimensional phase unwrapping using FFT’s. *IEEE Trans. Signal Process.*, 32(3):706–708, May 1994.

- [11] S. Rangan. Generalized approximate message passing for estimation with random linear mixing. arXiv:1010.5141v1 [cs.IT]., October 2010.
- [12] S. Rangan, A. K. Fletcher, V. K. Goyal, and P. Schniter. Hybrid approximate message passing with applications to structured sparsity. arXiv:1111.2581 [cs.IT], November 2011.
- [13] B. Schwarz. LIDAR: Mapping the world in 3D. *Nature Photonics*, 4(7):429–430, July 2010.
- [14] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu. On advances in statistical modeling of natural images. *J. Math. Imaging Vision*, 18(1):17–33, January 2003.
- [15] M. A. Stephens. Random walk on a circle. *Biometrika*, 50(3–4):385–390, 1963.
- [16] Elena Stoykova, A. Aydin Alatan, Philip Benzie, Nikos Grammalidis, Sotiris Malassiotis, Joern Ostermann, Sergej Piekh, Ventseslav Sainov, Christian Theobalt, Thangavel Thevar, and Xenophon Zabulis. 3-D time-varying scene capture technologies—A survey. *IEEE Trans. Circuits Syst. Video Technol.*, 17(11):1568–1586, November 2007.
- [17] M. Vetterli, P. Marziliano, and T. Blu. Sampling signals with finite rate of innovation. *IEEE Trans. Signal Process.*, 50(6):1417–1428, June 2002.
- [18] J. P. Vila and P. Schniter. Expectation-maximization Gaussian-mixture approximate message passing. arXiv:1207.3107 [cs.IT], July 2012.