Okinawa Institute of Science and Technology
Graduate University

Thesis submitted for the degree

Doctor of Philosophy

Algorithmic and Architectural Developments for Cryo-Electron Tomography

by

Faisal Mahmood

Supervisor: **Ulf Skoglund**
Co-Supervisor: **Hiroaki Kitano**

March, 2017

Declaration of Original and Sole Authorship

I, Faisal Mahmood, declare that this thesis entitled *Algorithmic and Architectural Developments for Cryo-Electron Tomography* and the data presented in it are original and my own work.

I confirm that:

- This work was done solely while a candidate for the research degree at the Okinawa Institute of Science and Technology Graduate University, Japan.
- No part of this work has previously been submitted for a degree at this or any other university.
- References to the work of others have been clearly attributed. Quotations from the work of others have been clearly indicated, and attributed to them.
- In cases where others have contributed to part of this work, such contribution has been clearly acknowledged and distinguished from my own work.
- None of this work has been previously published elsewhere, with the exception of the following:

Publications:

- 1) F.Mahmood, L.G. Öfverstedt, M. Toots, G. Wilken, U.Skoglund, "An Extended Field-based Method for Noise Removal from Electron Tomography Reconstructions" Nature Scientific Reports (Submitted).
- 2) F. Mahmood, M. Toots, L.G. Öfverstedt, U.Skoglund, "Algorithm and Architecture Optimization for 2D Discrete Fourier Transforms with Simultaneous Edge Artifact Removal" IEEE Access (Submitted).
- 3) F. Mahmood, N. Shahid, U. Skoglund, P. Vandergheynst, "Adaptive Graph-based Total Variation for Tomographic Reconstructions" IEEE Signal Processing Letters (Submitted).
- 4) F. Mahmood, N. Shahid, P. Vandergheynst U. Skoglund, "Graph-based Sinogram denoising for Tomographic Reconstructions" 38th IEEE Engineering in Medical & Biology Conference (EMBC) 2016, Orlando, FL.

- 5) F. Mahmood, M. Toots, U. Skoglund et al., “2D Discrete Fourier Transform With Simultaneous Edge Artifact Removal For Real-Time Applications” IEEE Field Programmable Technology (FPT) 2015 Queenstown, New Zealand.

Patents:

- 1) F. Mahmood, L.G. Öfverstedt & U. Skoglund, “Extended Field Iterative Reconstruction Technique (EFIRT) for Correlated Noise Removal from 3D Reconstructions” US Application Number: **US 14/770,245** PCT Application Number: **PCT/JP2014/001214**, Status/JP: Granted

- 2) F. Mahmood, M. Toots, L.G. Öfverstedt & U. Skoglund, “2D Discrete Fourier Transform With Simultaneous Edge Artifact Removal For Real-Time Applications”, PCT Application Number: **PCT/JP2016/003401**.

Date: March, 2017
Signature:

Abstract

Algorithmic and Architectural Developments for Cryo-Electron Tomography

Molecular structure determination is important for understanding functionalities and dynamics of macromolecules, such as proteins and nucleic acids. Cryo-electron tomography (Cryo-ET) is a technique that can be used to determine structures of individual macromolecules, thus providing snapshots of their native conformations. Such 3D reconstructions encounter several types of imperfections due to missing, corrupted and low-contrast data. This thesis focuses on the algorithmic and architectural aspects of improving and accelerating tomographic reconstructions specifically for Cryo-ET. The thesis explores modern compressed sensing and graph-based non-local approaches for noise removal and for partially recovering the missing wedge. These methods act as a proof-of-concept for the applicability of sparsity exploiting methods to tomographic image reconstruction. The thesis also explores, analyses and explains an extended field (EF)-based noise removal method. When used in conjunction with a variety of reconstruction procedures with a regularization capability it proved to be computationally efficient, reliable and stable. Through extensive empirical simulations it was shown that extending the reconstruction space reduces the error at a relatively lower regularization parameter thus allowing a better fit with the projections and preventing oversmoothing. Computational constraints are a major issue in speeding up tomographic reconstruction and refinement. One of the fundamental components, which often becomes a bottleneck in a variety of analytical tomographic reconstruction methods, is the 2D fast Fourier transform (FFT). Generally, 2D FFTs suffer from edge artifacts or series termination errors, which stem from the fact that two opposing edges of an image are often not periodic. These artifacts can propagate to next stages of processing and appear as errors in reconstructions. This thesis also explores simultaneous 2D FFTs and edge artifact removal for real-time applications. This was accomplished on a multi-FPGA (Field Programmable Gate Array) reconfigurable computing system with a high-speed bus. The algorithmic optimization and architecture are general and can be replicated to a variety of different hardware setups.

Acknowledgment

I wish to express my gratitude to a number of people, without whom this thesis would not have been possible.

First, I would like to thank my supervisor Prof. Ulf Skoglund for introducing me to the field of Cryo-Electron Tomography and all the interdisciplinary challenges associated with the field. For endless inspiring discussions, mentoring and for steering me in the right direction whenever it was needed. For teaching me the value of my work, for encouraging me to be independent by establishing a continuous improvement strategy. He taught me endurance and to keep a positive attitude which was needed for various parts of this work. He continues to provide inspiration and it has been a true honor to be his student over the past five years.

I would like to thank Prof. Hiroaki Kitano for his input, periodic meetings and continued support during the project. Prof. Gail Tripp for her continuous encouragement and support as my academic mentor.

I am also grateful to Nauman Shahid and Prof. Pierre Vandergheynst of EPFL, Switzerland for long discussions and for providing optimization tools for Graph-based total variation methods. I would also like to thank the faculty and staff at the Electrical Engineering Department at Imperial College London for hosting me as a visiting student (2013-14). Kuniyo Sueyoshi, Zhe Zhou, Sohaib Hussain and other graduate students at the Imperial also deserve to be thanked for making my stay in London memorable. I would also like to thank Märt Toots for suggesting software-based optimizations of periodic plus smooth decomposition. Dr. Lars-Göran Öfverstedt for teaching me how to use the tomography software package. Dr. Gunnar Wilken for mathematical assistance and suggesting experiments for extended field iterative reconstruction technique. Shizuka Kuda for logistical arrangements, doing all the paper-work for the equipment purchased and arranging my travels to various conferences, meetings, courses and trainings. I would also like to thank the OIST Graduate School and Student Support staff for supporting me over the past five years.

I would also like to thank the technical support and sales staff at National Instruments Research Austin, National Instruments Research Dresden and National Instruments Japan for endless discussions, trainings and application assistance on using their hardware systems. Dr. Steven D. Aird for providing language assistance for various parts of this thesis. I would like to thank my wife Zahra Noor for accompanying me during part of my stay in Japan and to the Japanese Government for funding my PhD at OIST. Finally, I would also like to thank Prof. Abraham J. Koster and Prof. Donald Bailey for suggesting revisions and corrections in the thesis.

Abbreviations

2D	Two Dimensional
3D	Three Dimensional
Å	Ångström
ARMs	Algebraic Reconstruction Methods
ART	Algebraic Reconstruction Technique
ASIC	Application Specific Integrated Circuit
AGTV	Adaptive Graph Total Variation
BRAM	Block Random Access Memory
CCD	Charge Coupled Detector
Cryo-EM	Cryo-Electron Microscopy
Cryo-ET	Cryo-Electron Tomography
CTF	Contrast Transfer Function
CS	Compressed Sensing or Compressive Sensing
CT	Computed Tomography
CSGT	Compressed Sensing Graph Total Variation
CSTV	Compressed Sensing Total Variation
DROP	Diagonally Relaxed Orthogonal Projections
DRAM	Dynamic Random Access Memory
DCT	Discrete Cosine Transform
EM	Electron Microscopy
EF	Extended Field
EFIRT	Extended Field Iterative Reconstruction Technique
FBP	Filtered Back-projection
FFT	Fast Fourier Transform
FRC	Fourier Ring Correlation
FSC	Fourier Shell Correlation
FT	Fourier Transform
FPGA	Field Programmable Gate Arrays
I/O	Input/Output
ROI	Region of Interest
RAPS	Radially Averaged Power Spectrum
SART	Simultaneous Algebraic Reconstruction Techniques
SER	Sparsity Exploiting Reconstruction
SEIR	Sparsity Exploiting Image Reconstruction
SEMs	Sparsity Exploiting Methods
SIRT	Simultaneous Iterative Reconstruction Technique

SNR	Signal to Noise Ratio
SPEM	Single Particle Electron Microscopy
SC	Stopping Criteria
TEM	Transmission Electron Microscope
Tk.	Tikhonov Regularization
RCD	Row and Column Decomposition
RAPS	Radially Averaged Power Spectrum
PSD	Periodic Plus Smooth Decomposition
PSF	Point Spread Function
PCI	Peripheral Component Interconnect
PCIe	PCI Express
PLDs	Programmable Logic Devices
PXIe	PCI for Industry Express
NI	National Instruments
NLTV	Non-Local Total Variation
NLM	Non-Local Means
NUFFT	Non-Uniform FFT
FlexRIO	Flexible Reconfigurable Input Output
OPSD	Optimized Periodic Plus Smooth Decomposition
TEM	Transmission Electron Microscopy
TV	Total Variation
TH	Tile Hopping
XTAL	X-Ray Crystallography

for mom and dad!

Summary of Contributions

Cryo-Electron Tomography (Cryo-ET) is a modern structure determination method that can preserve the conformations of molecules being imaged in their native states. Reconstructing Cryo-ET data is an *ill-posed inverse problem* due to missing, noisy and incomplete data, which leads to non-unique unstable solutions. ***This thesis focuses on developing and analyzing tomographic image reconstruction algorithms for denoising Cryo-ET reconstructions as well as developing architectures that could potentially accelerate reconstruction methods.*** The reconstruction and denoising problem is approached purely from an image reconstruction and algorithm development point of view. An analysis of the proposed methods for specific biological molecules is beyond the scope of this engineering oriented thesis. Architectural developments associated with FPGA-based reconfigurable computing methods are general and can be adapted to a variety of different applications. The following original contributions have been made in this thesis:

From an Image Reconstruction Perspective:

- The proposition of graph-based non-local sinogram denoising, its analysis and effectiveness for enhancing standard tomographic reconstruction methods. This method exploits the fact that the sinogram has structure since it is composed of projections from the same sample. Non-local processing methods can be used to denoise the sinogram based on similar pixels which are spatially far from each other (Chapter 2).

- The proposition of adaptive graph-based total variation as a compressed sensing-type sparsity exploiting image reconstruction method which can simultaneously reconstruct and denoise tomographic data. The study of the ability of the method to reconstruct from missing and noisy data. A unique aspect of the method is that it can be seen as a generalization of compressed sensing and total variation-type methods being extensively used, a fact of significance and interest for the image reconstruction community (Chapter 3).

- Analysis of an extended field-based reconstruction method that can be used to denoise reconstruction without direct manipulation of individual voxels or pixels and is computationally efficient and stable. Major contribution in this area was to enable the use of extended field with a wide variety of methods and to study its properties pertaining to the extension size and its ability to reduce the regularization parameter thus preserving the faithfulness of the reconstruction with the projection data (Chapter 5).

From a Reconfigurable Computing Perspective:

- Development of an FPGA-based 2D FFT architecture with simultaneous edge artifact removal for high-performance applications. This was achieved by making the following contributions:
- Optimizing the existing periodic plus smooth-based edge artifact removal scheme to reduce the access of DRAM and 1D FFT invocations for an efficient FPGA implementation.
- Designing a custom memory controller to overcome the 'memory wall' issue which results in reduced memory access speeds while accessing data stored column-wise in external memory.

A more detailed summary of contributions has been given at the end of Chapter 1.

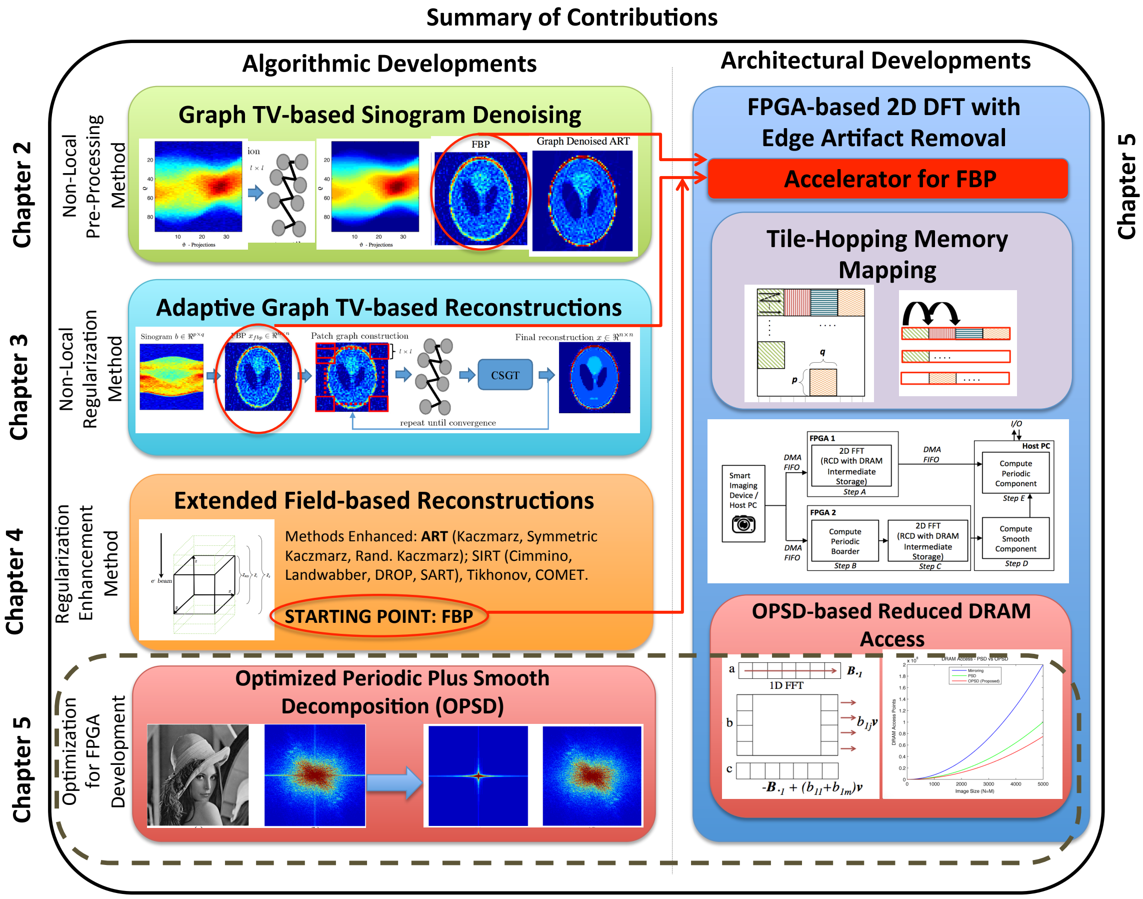

Summary of Contributions
Algorithmic Developments
Architectural Developments
Chapter 2
Non-Local Pre-Processing Method
Graph TV-based Sinogram Denoising
FBP
Graph Denoised ART
Chapter 3
Non-Local Regularization Method
Adaptive Graph TV-based Reconstructions
Patch graph construction
CSGT
Final reconstruction
repeat until convergence
Chapter 4
Regularization Enhancement Method
Extended Field-based Reconstructions
Methods Enhanced: ART (Kaczmarz, Symmetric Kaczmarz, Rand. Kaczmarz); SIRT (Cimmino, Landwabber, DROP, SART), Tikhonov, COMET.
STARTING POINT: FBP
Chapter 5
Optimization for FPGA Development
Optimized Periodic Plus Smooth Decomposition (OPSD)
Chapter 5
FPGA-based 2D DFT with Edge Artifact Removal
Accelerator for FBP
Tile-Hopping Memory Mapping
p
q
Smart Imaging Device / Host PC
DMA FIFO
FPGA 1
2D FFT (RCD with DRAM Intermediate Storage)
Step A
DMA FIFO
I/O
Host PC
Compute Periodic Component
Step E
FPGA 2
Compute Periodic Boarder
Step B
2D FFT (RCD with DRAM Intermediate Storage)
Step C
Compute Smooth Component
Step D
OPSD-based Reduced DRAM Access
a
b
c
1D FFT
$B_{\cdot 1}$
$b_{1j}v$
$-B_{\cdot 1} + (b_{11}+b_{1m})v$
DRAM Access - PSD vs OPSD
DRAM Access Points
Image Size (N=M)

Figure 1: A graphical representation of algorithmic and architectural contributions.

Contents

Declaration of Original and Sole Authorship iii

Abstract v

Acknowledgment vii

Abbreviations ix

Summary of Contributions xiii

Contents xv

List of Figures xix

List of Tables xxvii

1 Introduction 1
- 1.1 Chapter Outline . . . 1
- 1.2 Macromolecular Structure Determination . . . 1
- 1.3 Cryo-Electron Tomography (Cryo-ET) . . . 2
  - 1.3.1 Cryo-ET: Methodology . . . 4
- 1.4 Cryo-ET: Noise, Imperfections and Challenges . . . 6
- 1.5 Tomographic Image Reconstruction . . . 7
  - 1.5.1 Analytical Reconstruction Methods . . . 7
  - 1.5.2 Iterative Reconstruction Methods . . . 12
- 1.6 High-Performance and Reconfigurable Computing . . . 15
- 1.7 Contributions . . . 16

2 Graph-based Non-Local Sinogram Denoising 19
- 2.1 Chapter Outline . . . 19
- 2.2 Sinogram Denoising . . . 19
- 2.3 Introduction to Graphs . . . 22
- 2.4 Graph-based Sinogram Denoising . . . 24
  - 2.4.1 Graph Construction for Sinogram Denoising . . . 25
  - 2.4.2 Graph Total Variation Denoising . . . 26
- 2.5 Reconstruction Phase . . . 27

2.6 Experiments and Results . . . 28

2.6.1 Simulation Setup . . . 28

2.7 Limitations and Shortcomings . . . 40

2.8 Conclusions . . . 41

3 Adaptive Graph Total Variation for Tomographic Reconstructions 43

3.1 Chapter Outline . . . 43

3.2 The Story So Far . . . 43

3.3 Sparsity Exploiting Image Reconstruction (SEIR) . . . 44

3.4 Compressed Sensing & Adaptive Graph Total Variation . . . 46

3.4.1 Optimization Problem . . . 46

3.4.2 Compressed Sensing . . . 46

3.4.3 Graph Total Variation . . . 48

3.4.4 Graph Construction for Total Variation . . . 49

3.4.5 Adaptive Graph Total Variation Regularization . . . 50

3.5 Optimization Solution . . . 50

3.5.1 Forward-Backward based Primal Dual Solver . . . 50

3.5.2 Algorithm . . . 51

3.5.3 Computational Complexity . . . 51

3.6 Working Explanation of AGTV . . . 53

3.7 Experimental Results . . . 54

3.8 Shortcomings & Limitations . . . 68

3.9 Conclusions . . . 69

4 Extended Field-based Noise Removal from Tomographic Reconstructions 71

4.1 Chapter Outline . . . 71

4.1.1 Introduction . . . 71

4.2 Preliminaries & Simulation Experiments . . . 73

4.2.1 Extended Reconstructions . . . 74

4.2.2 ART with Extended Field (EART) . . . 74

4.2.3 Simulations with EART . . . 77

4.2.4 SIRT with Extended Field (ESIRT) . . . 84

4.2.5 Simulations with ESIRT . . . 85

4.2.6 Simulations with Extended Tikhonov Regularization (ETR) . . 85

4.3 Quantifying the Amount of Noise Removed . . . 86

4.4 Enhanced Regularization Using Extended Field . . . 88

4.5 Optimal Extension Size . . . 89

4.6 Experimental Results from 3D Data . . . 91

4.7 Limitations and Shortcomings . . . 94

4.8 Conclusions . . . 94

5 Algorithm and Architecture Optimization for 2D Discrete Fourier Transforms with Simultaneous Edge Artifact Removal 97

5.1 Chapter Outline . . . 97

5.2 Accelerating Analytical Tomography . . . 98

5.3 Introduction to 2D FFTs 99

5.3.1 Contributions 100

5.4 Background 101

5.4.1 High Performance 2D FFTs using FPGAs 101

5.4.2 DRAM Intermediate Storage Problem 101

5.4.3 Edge Artifacts 104

5.4.4 LabView FPGA Semi-High Level Design Environment 105

5.5 Periodic Plus Smooth Decomposition (PSD) for Edge Artifact Removal 105

5.5.1 Profiling PSD for FPGA Implementation 108

5.6 Optimized Periodic Plus Smooth Decomposition (OPSD) 109

5.7 Tile-Hopping Memory Mapping for 2D FFTs 113

5.8 Experimental Results and Analysis 115

5.8.1 Hardware Configuration and Target Selection 115

5.8.2 Experimental Setup 116

5.8.3 Performance Evaluation 118

5.8.4 Energy Evaluation 120

5.9 Application: Filtered Back-projection for Tomography 122

5.10 Conclusion 123

6 Conclusions and Future Directions 125

6.1 Summary and Conclusions 125

6.2 Original Contributions 127

6.3 Future Directions 127

A Details Regarding Multimedia Files 129

Bibliography 131

List of Figures

1 A graphical representation of algorithmic and architectural contributions. xiv

1.1 The overall process of cryo-electron tomography from data collection (*forward problem*) to reconstruction and refinement (*inverse problem*). . 3

1.2 Figure representing data collection in Cryo-ET by rotating the sample at various angles. It should be noted that unlike most other biomedical imaging tomographic modalities Cryo-ET does not involve the rotation of the source and the detector rather the sample is rotated. Physical constraints limit the rotation of the sample to a certain angle. 5

1.3 Figure showing graphical representation of the radon transform. A single ray passing through an object $f(x, y)$, when several rays pass through the $f(x, y)$ they form a projection of the object at an angle θ with the x-axis. 9

1.4 Figure showing graphical representation of the sinogram for a small box and the notorious Shepp-Logan phantom. In this case the sinogram is built column-wise, *i.e.*, each column represents a distinct 1D projection from the 2D object at a particular angle. 9

1.5 Figure showing the graphical representation of the projection slice theorem or Fourier slice theorem. 11

1.6 FBP reconstructions (Linearly interpolated, 'Ram-Lak' filter) with changing number of projections. The streaking artifacts show for lower number of projections show that there is not enough information to confine certain parts of the projections. The projections were simulated without any noise. 12

1.7 FBP reconstructions (Linearly interpolated, 'Ram-Lak' filter) showing the effect of the missing wedge in the reconstructed phantom as well as the frequency domain. The projections were simulated without any noise. The 2D FFT shown is with the zero frequency component shifted to the center. Higher frequencies are attenuated due to the Ram-Lak filter. 13

1.8 A graphical representation of algorithmic and architectural contributions. 18

2.1 Various tomographic reconstruction improvement methods summarized. Most methods are based on local image processing paradigms whereby only proximal pixels or voxels are taken into account while processing the data. This chapter focuses on a non-local pre-processing method which is proceeded by a non-local sparsity exploiting regularization-based method in Chapter 3. 21

2.2 a) The Shepp Logan Phantom. b) Noisy sinogram of the Shepp Logan phantom (relative noise 0.08) c) Graph Denoised Sinogram. d) Smooth Phantom. e) Noisy sinogram of the Smooth phantom (relative noise 0.05) f) Graph Denoised Sinogram. Clearly, the projections have a smooth structure, which when exploited, denoises these sinograms significantly. 25

2.3 The overall process of Graph-based Sinogram denoising showing graph construction on the sinogram. 26

2.4 Recovered Shepp-Logan phantoms and intensity profiles for the following methods: 1) Filtered back projection (FBP), linearly interpolated, Ram-Lak filtered. 2) Graph denoised FBP (FBP-GD). 3) ART reconstruction (without denoising) and 4) Graph denoised ART / SIRT reconstruction (ART-GD, SIRT-GD). Images can be best viewed in color in the electronic version of the chapter. 31

2.5 ℓ_2 reconstruction error variations with k iterations, for the following methods: 1) Filtered back projection (FBP), linearly interpolated, ram-lak filtered. 2) Graph denoised FBP (FBP-GD). 3) ART reconstruction (without denoising) and 4) Graph denoised ART reconstruction (ART-GD, SIRT-GD). 31

2.6 ℓ_2 reconstruction error variations with k iterations, recovered phantoms and their intensity profiles for the following methods: 1) Filtered back projection (FBP) 2) Graph denoised FBP (FBP-GD) 3) ART / SIRT reconstruction (without denoising) and 4) Graph denoised ART / SIRT reconstruction (ART-GD, SIRT-GD). Images can be best viewed in color in the electronic version of the chapter. 33

2.7 The phantom, reconstructions and intensity profiles of reconstructing a 128×128 ring phantom from 71 equally spaced projections with a missing wedge ($-70^\circ \leq \theta \leq +70^\circ$) in the middle. It can be clearly seen that the effect of missing wedge elongation is prominent in the graph-denoised reconstructions. However, it can be seem from the intensity profiles that the overall noise in the reconstruction is much more constrained. 34

2.8 Radially averaged power spectrum (RAPS) of ring phantom reconstructions using FBP(linearly interpolated, ram-lak filtered), FBP using GD-Sinogram ART (Kaczmarz, SC=100), ART using GD-Sinogram, SIRT(Cimmino, SC=100) and SIRT using GD-Sinogram. This clearly shows that all methods are able to determine the basic structure of the phantom as seen from the lower wave number values. Due to the missing wedge and noise some methods add extra high resolution details to the result which are erroneous and incorrect. These erroneous high resolution details can be reduced by non-locally denoising the sinogram using the proposed method. 36

2.9 The phantom, reconstructions and intensity profiles of reconstructing a 256 × 256 Shepp-Logan phantom from 71 equally spaced projections with a missing wedge ($-70^\circ \leq \theta \leq +70^\circ$) in the middle. It can be clearly seen that the effect of missing wedge elongation is prominent in the graph-denoised reconstructions. However, it can be seen from the intensity profiles that the overall noise in the reconstruction is much more constrained. 38

2.10 RAPS of the Shepp-Logan phantom reconstructions using FBP(linearly interpolated, ram-lak filtered), FBP using GD-Sinogram ART (Kaczmarz, SC=100), ART using GD-Sinogram, SIRT(Cimmino, SC=100) and SIRT using GD-Sinogram. This clearly shows that all methods are able to determine the basic structure of the phantom as seen from the lower wave number values. Due to the missing wedge and noise some methods add extra high resolution details to the result which are erroneous and incorrect. These erroneous high resolution details can be reduced by non-locally denoising the sinogram using the proposed method. 39

2.11 Figure showing the FBP and GD-FBP reconstruction of Cryo-ET emulated data as well as their RAPS and intensity profiles. The data used is from EMD-4067. It can be seen that although the image reconstructed from the graph-denoised image is smoother it loses significant frequency details. There high frequency details could also be from noise. The image was reconstructed from 141 1D projections. Data used is from an Influenza virus available from opensource EM-Data-Base Entry: 4067. . 40

3.1 For the construction of patch graph, $x_{fbp} \in \mathbb{R}^{n \times n}$ is divide into n^2 overlapping patches of size $l \times l$ each. 49

3.2 The complete methodology for AGTV. The input sinogram / projections $b \in \mathbb{R}^{p \times q}$ is first used to obtained a filtered back projection (FBP) $x_{fbp} \in \mathbb{R}^{n \times n}$. It is then used to construct the initial patch graph $\mathcal{G}$ to be used by the CSGT method. The output of CSGT is used to refine / reconstruct the graph and this process is repeated until convergence. 50

3.3 A comparison of the Total Variation (TV) and Adaptive Graph Total Variation (AGTV) priors for the methods CSGT and AGTV. The TV prior does not connect patches ‘a’ and ‘b’ which possess structural similarity, whereas the GTV prior connects them because the $\mathcal{K}$-nearest neighbor graph is not restricted to spatial neighbors only. Furthermore, this connection keeps getting stronger due to iterative removal of noise and graph updates in every iteration. 53

3.4 Comparative analysis of reconstructions and intensity profiles of reconstructing a Shepp-Logan phantom using various reconstruction methods. 57

3.5 Comparative analysis of reconstructing a Torso phantom using various reconstruction methods. 59

3.6 Comparative analysis of reconstructing a Shepp-Logan phantom using various reconstruction methods at 5% and 10% and 20% Poisson noise. FBP (Linearly interpolated, Cropped Ram-Lak filter); ART (Kaczmarz/Randomized Kaczmarz, Relaxation Parameter (η) = 0.25, Prior: FBP, Stopping Criteria = 100 iterations); SIRT (Cimmino/SART, (η) = 0.25, Prior: FBP, Stopping Criteria = 100 iterations); CS (500 Iterations, Prior: FBP); CSTV ($\lambda = 0.5$, $\gamma = 0.1$, Prior: FBP, Stopping Criteria = 100 iterations); CSGT ($\lambda = 0.5$, $\gamma = 0.2$, Prior: Patch Graph from FBP, Stopping Criteria = 100 iterations); AGTV ($\lambda = 0.5$, $\gamma = 1$, Prior: Patch Graph from FBP updated every iteration, I and J in Algorithm 2 set to 30). . 60

3.7 Comparative analysis of reconstructions and intensity profiles in both directions of a Shepp-Logan phantom using various reconstruction methods at 5% noise. It can clearly be seen that AGTV behaves better than all other methods, although it does not fully reconstruct the missing wedge it is able to partly reconstruct it. 62

3.8 Fourier Ring Correlation of the original Shepp-Logan phantom with reconstructions from various reconstruction methods. Clearly there is hardly any loss of resolution in the case of AGTV, mild loss is expected since the method cannot perfectly reconstruct the missing wedge and all noise cannot be removed. 63

3.9 Comparative analysis of the radially averaged power spectrums of the Shepp Logan phantom reconstructed employing various reconstruction methods. It can be seen that AGTV is the closest to the spectrum of the Shepp-Logan ground truth. 64

3.10 2D FFT Magnitude components of the Shepp-Logan phantom, FBP reconstruction clearly showing the missing wedge and the AGTV reconstruction showing the extent to which the missing wedge can be recovered. The recovery is attributed to the CS-type nature of the sparsity exploiting image reconstruction method presented here. 64

3.11 A small experiment corresponding to the reconstruction of a 32 × 32 Shepp-Logan phantom from 36 projections $b \in \mathbb{R}^{36}$ using the pre-tuned parameters $\lambda = 0.1, \gamma = 5$ for different values of $\mathcal{K}$ ranging from 5 to 50. The results clearly show that the reconstruction is quite robust to the choice of $\mathcal{K}$, with a small error variation. 66

3.12 A small experiment corresponding to the reconstruction of a 32×32 Shepp-Logan phantom from 36 projections $b \in \mathbb{R}^{36}$ using the full parameter gird $\lambda = (0.1, 1), \gamma = (0.1, 10)$ for a fixed value of $\mathcal{K} = 15$. The minimum clustering error occurs at $\lambda = 0.2, \gamma = 0.1$. The results clearly show that the reconstruction error increases smoothly with the parameters. 66

3.13 Figure showing results from FBP, AGTV reconstruction of Cryo-ET emulated data as well as their RAPS and intensity profiles. Data used is from an Influenza virus available from opensourse EM-Data-Base Entry: 4067. The image was reconstructed from 141 1D projections. As compared to sinogram denoising AGTV preserves more high frequency details. 67

3.14 Figure showing run-time for AGTV corresponding to 2D tomography problems reconstructed from 1D data on an Intel Xeion Machine (E5450 12M Cache, 3.00 GHz, 128GB RAM). 69

4.1 a) Consistent System of Equations having an exact solution, showing how Kaczmarz method approaches the solution starting from an initial guess. b) In case of an inconsistent noisy system the solution would be in a region bound by hyperplanes rather than at their intersection. . . 74

4.2 Reconstruction of a binary phantom with an extra region larger than the ROI. Unconstrained noise is redistributed out of the geometrically constrained ROI. The extra region gives flexibility to minimize data errors and maximize consistency of the solution within the ROI. 75

4.3 Image showing how the sinogram is extended to achieve an extra region outside the ROI. If a sinogram corresponding to a $N_1 \times M_1$ phantom is to be extended to a reconstruction corresponding to $N_2 \times M_2$, the sinogram is extended by zero-padding each projection with $(\sqrt{M_2^2 + N_2^2} - \sqrt{M_1^2 + N_1^2})/2$ on each side. For example, if the sinogram of a 64×64 phantom is extended to a 128×128 phantom the extension on each side of the sinogram would be 45. 76

4.4 Flow diagram showing the simulation setup for various extended field simulations with several different iterative reconstruction methods. It should be noted here that in step 3 the sinogram is zero-padded in the direction of the projections *i.e.* adding an extra region to each projection. 78

4.5 Comparative analysis of reconstructing a binary (relative noise = 0.05) and Shepp-Logan (relative noise = 0.1) phantom with ART and EART shows that EART can achieve better results with a faster semi-convergence rate and lower error rate than ART. Codes to generate these figures are available with this thesis. 79

4.6 a-f)Intensity profiles of a line through the original and reconstructed binary and Shepp-Logan phantoms using FBP, ART (Kaczmarz) and EART (Extended Kaczmarz). The EART reconstructions show a relatively better fit to the original phantom as compared to EART reconstructions. 80

4.7 Comparative analysis of ART and EART with respect to changing relaxation parameter γ. 80

4.8 Comparative analysis of reconstructing a smooth phantom (relative noise = 0.05) and Shepp-Logan (relative noise = 0.1) with SIRT and ESIRT shows that ESIRT can achieve better results with a faster convergence and lower error rate as compared to SIRT. 81

4.9 Intensity profiles of a line through the original and reconstructed smooth and Shepp-Logan phantoms using FBP, SIRT (Cimmino) and ESIRT (Extended Cimmino). The ESIRT reconstructions show a better fit to the original phantom as compared to SIRT reconstructions. 82

4.10 Comparative analysis of reconstructing a binary (relative noise = 0.05) and Shepp-Logan (relative noise = 0.1) phantom with Tikhonov and extended Tikhonov regularization. Extended Tikhonov performs better than regular Tikhonov regularization. Intensity profiles of for both phantoms show that extended Tikhonov performs better than Tikhonov and has a higher regularization effect. 83

4.11 Experimental design to quantify the noise in the extended field. The pixels in the reconstructed ROI are set to zero and projections are taken relative to the ROI and compared with noise added to individual projections of the initial sinogram. 86

4.12 Comparison of noise removed, *i.e.*, noise redistributed into the extended field and the noise originally added to the projection at 25^o, 50^o, 75^o and 100^o for extended ART, SIRT and Tikhonov reconstructions of a Shepp-Logan phantom. The good fit shows that the amount of noise removed by the flexibility provided by the extended field is approximately equivalent to the amount of noise added while conducting the simulation (Table 4.4). Results for the entire sinogram and similar results for ART and SIRT have been presented in the supplementary material. 87

4.13 Comparative analysis of error rates for Tikhonov and Extended Tikhonov reconstructions at a varying regularization parameter ($0.1 \leq \lambda \leq 3$). The error rates are calculated corresponding to the ROI for Extended Tikhonov reconstructions. It can be clearly seen that Extended Tikhonov achieves a much better error rate at a lower λ than regular Tikhonov. . 88

4.14 Comparative analysis of error rates for ART, SIRT and Tikhonov reconstructions at varying extension steps $0, 2, 4, 8, ...512$ starting from a 64×64 phantom. It can be clearly seen that most of the benefit of an extended field is achieved if the size of the extension is half the actual reconstruction size. All reconstructions show a sharp improvement during the first few small extension steps followed by a saturation of error improvement for larger extensions. 90

4.15 Figure showing successive reconstructions on a large volume by increasing in z-direction (*i.e.*, reconstructions with an extended field). The volume may also be increased by increasing the reconstruction space in y-direction. 92

4.16 Extracted ROIs (355 × 355 × 255) of successive reconstructions of a colloidal silica sample with increasing z. Reconstructions with an extended field (*i.e.* $R_2 - R_8$) have a lower noise within the ROI. The map grid size for one pixel is 5.626Å and the average size of a colloidal silica nanoparticle is 31.5nm. The large particle in the center of the reconstruction has lower density as compared to the smaller particles because it is water in the form of crystallized ice. Videos can be viewed in the multimedia supplement with this thesis. 93

4.17 Intensity profile of a single line through a slice of the 3D reconstruction at two different values of z. The peak represents the signal. For a reconstruction with an extended field the signal within the ROI is significantly stronger as compared to the noise. 94

5.1 Image showing the ROI reconstruction problem in tomography. Microscopes these days produce large size micrographs (usually 4096 × 4096). However, computational constraints don't allow for such large reconstructions to be conducted. Usually the projections are cropped to extract a region of the aligned tomogram which is then reconstructed. This leads to information in the extracted projections which does not belong to the ROI. 99

5.2 a) An overview of Row-Column Decomposition (RCD) for 2D FFT implementation. Intermediate Storage is required because all elements of the row-by-row operations must be available for column-by-column processing. b) An overview of strided Column-wise access from DRAM as compared to trivial row-wise access. An entire row of elements must be read into the row buffer even to access a single element within a specific row. 102

5.3 a) An overview of the DRAM hierarchy. b) Image showing the structure of a single DRAM bank. c) Flow chart explaining additional latency introduced when a new row has to be referred into the row buffer to access a specific element. 103

5.4 1a) An image with non-periodic boundary. 1b) 2D DFT of 1a. 1c) DFT of the Smooth Component ,*i.e.*, the removed artifacts from 1a. 1d) Periodic Component i.e DFT of 1a with Edge Artifacts removed. 1e) Reconstructed Smooth Component. 1f) Reconstructed Periodic Component. 107

5.5 A top-level architecture for OPSD using two FPGAs and a host PC connected over a high-bandwidth bus. The steps are associated with algorithm 5. 109

5.6 Graph showing DRAM access (equal to number of DFT points to be computed) with increasing image size for Mirroring, Periodic Plus Smooth Decomposition (PSD) and our proposed Optimized Period Plus Smooth Decomposition (OPSD). 112

5.7 Image showing tile-hopping. a) Image-level view showing tiles. b) DRAM-level view showing tile placement while writing. c) DRAM-level view showing column reading from the tiles. 114

5.8 Block diagram of a PXIe based multi-FPGA system with a host PC controller connected through a high-speed bus on a PXIe chassis. . . . 116

5.9 Functional block diagram of PXIe based 2D FFT implementation with simultaneous edge artifact removal using optimized periodic plus smooth decomposition. The OPSD algorithm is split among two NI-7976R (Kintex-7) FPGA boards with 2GB external memory and a host PC connected over a high-bandwidth bus. The image is streamed from the PC controller to FPGA 1 and FPGA 2. FPGA 1 calculates the row-by-row 1D FFT followed by column-by-column 1D FFT with intermediate tile-hopping memory mapping and sends the result back to the host PC. FPGA 2 receives the image, calculates the boundary image and proceeds to calculate the 1D FFT column-by-column FFT using the shortcut presented in Equation. 16 followed by row-by-row 1D FFTs and the result is sent back to the host PC. 117

5.10 Block diagram of 2D FFT showing data transfer between external memory and local memory scheduled via a Control Unit (CU). 118

5.11 Performance evaluation in terms of frames per second for a) 2D FFTs with Tile-Hopping Memory Pattern. b) 2D FFTs with Edge Artifact Removal (EAR) using OPSD. The performance evaluation shows the significance of the two optimizations proposed. Both axes are on a log scale. 119

5.12 Figure showing a thin slice of filtered back-projection results by reconstructing a $128 \times 128 \times 128$ Shepp-Logan phantom. The 3D density was reconstructed from 180 equally spaced simulated projections using standard linearly interpolated FBP and using a ram-lak filter. It can be seen that the results from the FPGA+CPU solution have some errors this is due to the fact that the 2D FFT of each projection is less accurate. The results are good enough to be used as a prior for further optimization based refinement methods. 122

List of Tables

2.1 Comparison of Regular and Graph Denoised (GD) Reconstructions . . 30

4.1 Comparison of error rates for ART and EART using various phantoms at 20% noise 80
4.2 Comparison of error rates for SIRT and ESIRT using various phantoms at 20% noise 82
4.3 Comparative analysis of error rates for Tikhonov and extended Tikhonov reconstructions at varying relative noise and number of projections. . . 82

5.1 Comparing Mirroring, PSD and OPSD 111
5.2 Comparison of OPSD 2D FFT with regular RCD-based implementations 120
5.3 DRAM Energy Consumption Baseline vs Tile-Hopping 121
5.4 2D FFT + EAR Energy Consumption Baseline vs Optimized (OPSD+Tile Hopping) 121
5.5 Comparing Filtered Back-projection Runtime 123

Chapter 1

Introduction

1.1 Chapter Outline

In this chapter we introduce the field of Cryo-Electron Tomography (Cryo-ET) as a structure determination method and discuss its benefits and limitations. We also review various analytical and iterative tomographic reconstruction methods as well as the fundamental mathematical concepts associated with these methods such as Radon transform and the Fourier slice theorem. All tomographic reconstruction methods have a high computational cost specifically for large size reconstructions. Analytical methods can provide a primitive reconstruction which can then be used as a starting point (prior) for more complex iterative methods and help speed up convergence. Although, analytical methods such as FBP perform relatively faster than iterative methods, they become increasingly complex for large image sizes. Reconfigurable computing can be used as an acceleration platform for speeding-up some of these methods and will also be reviewed in this chapter. Finally, we present the algorithmic and architectural contributions of this thesis.

1.2 Macromolecular Structure Determination

Macromolecules such as proteins, amino acids, and nucleic acids are important biological molecules, which possess important functional information within their structure. Structure determination has various applications, specifically for understanding interactions between proteins which is of vital importance for drug development [1]. Most basic biological questions of significance can be tracked back to the cell. Obtaining high resolution images of macromolecular complexes and their interactions with the cell is of immense interest to the biological and medical community. The resolution of modern light microscopes, that can examine the cell *in vivo*, is limited by the wavelength of visible light (400 - 700 nanometers). There are three main physical techniques for three dimensional (3D) structure determination of biological macromolecules, x-ray crystallography (XTAL), nuclear magnetic resonance (NMR) and single particle electron microscopy (SPEM). Despite the fact that x-ray crystallography can achieve high resolution and is able to determine a vast range of structures from very small to large, crystallization and averaging are its limiting factors. Approximately only 10% of the known proteins can be crystallized, other proteins need to be fragmented before crystallization, making the overall process tedious. NMR on the other hand has a size limitation, the larger the size of the protein the more difficult it is to apply NMR. Structures of proteins up to 20-25kDa are easily determined after which it becomes increasingly difficult [2]. That said, NMR is a relatively new technique and its capability to determine structures of higher sized proteins is continuously increasing.

Cryo-Electron Tomography (Cryo-ET) although having a lower resolution has the ability to preserve the native structure of the macromolecules due to rapid cooling of the sample to a lower temperature. Moreover, Cryo-ET requires much lower amount of sample as compared to the other two techniques. This can be quite beneficial for eukaryotic proteins as they are difficult to purify in large quantity and are tedious for XTAL or NMR [3].

Drawback of Averaging Methods

Proteins are flexible since they do not have a rigid structure. Averaging techniques lose spatio-temporal information regarding the conformation or flexibility of the protein structure. During averaging a huge number of similar particles are averaged to form the final structure. During this process relatively rigid segments become prominent and significant information about the conformation of the molecule is lost. For example during x-ray crystallography flexible segments often do not have a well defined electron density, and thus averaging leads to the loss of information about the flexibility.

Single Particle Cryo-Electron Microscopy (SPEM)

Single particle Cryo-EM technique involves averaging of molecular structures by determining and superimposing macromolecules in different angular orientations either from single or multiple micrographs to reconstruct a 3D structure. Translation and in-plane rotation along with cross-correlation are used for determining similar molecules which are then separated into classes according to their angular orientations. SPEM requires less amount of sample and the sample is rapidly frozen at low temperature. One of the problems with this technique is that it is difficult to classify molecules which have very little difference or have a very close projection angle, making it difficult to use single particle technique with molecules smaller than 200kDa because there are issues with assignment of projection angles. Similar to other averaging method,s single particle technique also loses information about the flexibility of molecules. Still, single particle cryo-EM has been used to determine large macromolecular structures like viruses and the ribosome at high resolution [4].

1.3 Cryo-Electron Tomography (Cryo-ET)

Cryo-electron tomography (Cryo-ET) is the only structure determination method that can determine 3D structures of individual biological molecules in their native states. It preserves the occurrence of many different molecular conformations due to rapid freezing at liquid nitrogen temperature. Since Cryo-ET does not necessarily require a lot of averaging [5, 6], it can provide vital information for examinations of proteins

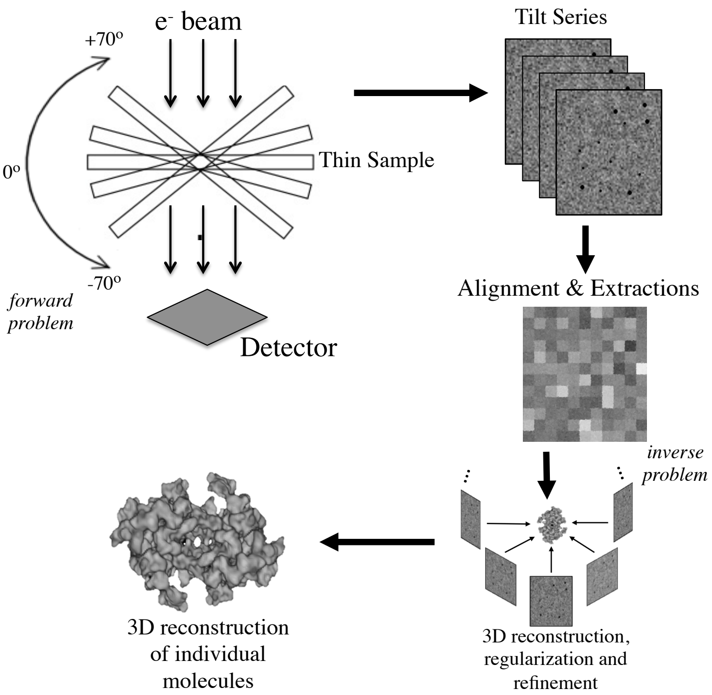


Figure 1.1: The overall process of cryo-electron tomography from data collection (*forward problem*) to reconstruction and refinement (*inverse problem*).

in situ, interactions among proteins, and analysis of molecular dynamics. Cryo-ET involves a transmission electron microscope (TEM) to take images of a cryo-specimen at successive angles by tilting the specimen in the electron microscope. These 2D images, often referred to as *tilt series*, are then aligned, and the desired region of reconstruction is extracted from each image. The 2D aligned extractions are then used to reconstruct a 3D tomogram via filtered back-projection (FBP) or other reconstruction and refinement methods [7] (Figure 1.1).

Cryo-ET has the same advantages of rapid freezing and low sample requirement as single particle cryo-EM but it has two additional advantages:

1. It does not necessarily involve averaging which gives us the ability to study flexible multi-domain proteins [5]. However, subtomogram averaging may be required to achieve higher resolution structures. During sub-tomogram averaging one may choose what to average, *i.e*, averaging of shape classified regions.

2. We do not have to determine and classify a lot of molecules according to their angular orientations, which gives us the ability to study relatively smaller molecules.

The drawbacks of Cryo-ET involve lower resolution compared to other methods. However, as mentioned using sub-tomogram averaging can lead to high resolution structures from Cryo-ET [8, 9]. The theoretical resolution is generally limited by the electron dose from the microscope, the thickness of the specimen (D) and the number of tilts (n, calculated over 180°) and is given by the Crowther criterion [10],

$$R = \frac{\pi D}{n}. \tag{1.1}$$

A lower dose is generally used to prevent degradation of the sample, this results in a lower contrast electron micrograph so a compromise must be made to obtain an effective image. For example on a cryo-biological sample which is being imaged by a 300keV TEM a dose of $40e/\text{Å}^2$ is often considered to be optimal. TEM Specimen thickness is determined by the size of the molecule under study; larger molecules require thicker specimens and have lower resolution.

1.3.1 Cryo-ET: Methodology

Cryo-ET generally involves the following steps:

1. Sample Preparation and Imaging.
2. Alignment and 3D Reconstruction.
3. Refinement and Noise Removal.

Sample Preparation

Cryo-ET samples are prepared and subsequently flash-frozen at liquid ethane temperature (-180°) at an extremely fast speed ($\sim 10^{-4}$ sec), which keeps the conformation of the molecules intact and prevents water molecules from crystallization. The thickness of the sample is preferred to be as small as possible, although as mentioned earlier it is limited by the size of the molecule. As we do not need crystals for cryo-specimen preparation we have relatively more freedom to choose the solution conditions. The sample is placed in a thin electron microscope (EM) grid and is rapidly submerged into liquid ethane.

Image Acquisition

After the sample is prepared it is exposed to vacuum and is then bombarded with a controlled dose of electrons in a transmission electron microscope (TEM). Interpreting the scattering of electrons forms a 2D image, the TEM has magnetic lenses to magnify and focus the image on a Charged Coupled Device (CCD) or a modern direct electron CMOS detector by which it is digitized. The specimen is tilted up to a range of angles in both directions with a constant step size to obtain multiple 2D projections of the cryospeimen often referred to as *micrographs*. Although the tilting process is fine, it is subject to mechanical errors and the specimen has to be refocused for every tilt. Simulating the image acquisition is commonly referred to as the *forward problem*, a detailed account of this has been presented in [11, 12].

Tilt-Series Alignment

Accurate alignment of the series of 2D images at different tilt angles is very important for getting a good 3D map. There are two commonly used methods for alignment

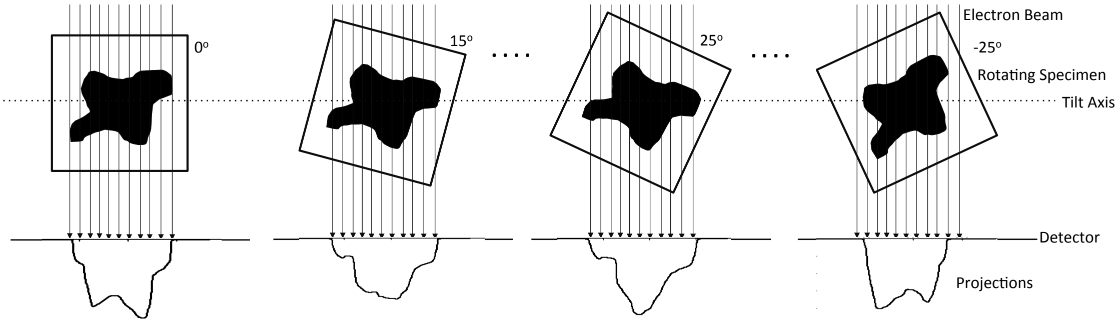


Figure 1.2: Figure representing data collection in Cryo-ET by rotating the sample at various angles. It should be noted that unlike most other biomedical imaging tomographic modalities Cryo-ET does not involve the rotation of the source and the detector rather the sample is rotated. Physical constraints limit the rotation of the sample to a certain angle.

a) fiducial markers (gold markers) b) sequential cross-correlation. Spherical gold particles having a diameter of ∼10nm can be used as markers which appear in all 2D images because they are electron dense and create high-contrast circle like features. After the dataset is recorded the gold-markers can be identified and 'picked' (localized) either manually or automatically. Once the markers are localized, least squares tracking of these can be used to determine the relative shift and the rotation of the electron micrograph. Alternatively, a cross-correlation technique can be used that tracks similar image features, which are present in all micrographs, and treats them as markers. Although automated methods based on cross-correlation and tracking features or patches have been developed such methods are highly sensitive to noise in low-dose micrographs, so using gold markers has proven to be more accurate [12]. The aligned images are then used for further processing stages.

Contrast Transfer Function (CTF) Deconvolution

The CTF is simply the Fourier transform of the point spread function. Generally a radial averaging of the CTF is used for analysis. Micrographs collected from TEMs are a projection of the specimen's electrostatic potential, convolved with the with the inverse Fourier transform of the CTF. The CTF oscillates around zero, this modulates the amplitude of the signal in Fourier space and also reverses the phase at some frequencies. This is not a problem if the expected resolution is below the first zero of the CTF. However, for higher resolution images the CTF must be corrected. Accurate CTF determination is not trivial since the SNR of the micrographs is quite low. More details regarding CTF estimation and deconvolution can be found in [12].

Tomographic Reconstruction

Once the collected tilt-series is aligned the 2D projections can be reconstructed to form a 3D density or tomogram. Fundamental tomographic reconstruction is based on the Fourier slice theorem also known as the projection slice theorem. However, due to the fact that Cryo-ET is a non-trivial reconstruction problem affected by issues discussed in 1.4. A more detailed review of tomographic image reconstructions has been presented in section 1.5.

1.4 Cryo-ET: Noise, Imperfections and Challenges

Equation 1.1 gives the maximum theoretical resolution possible. However this is never achievable since 3D reconstruction from TEM images is a severely *ill-posed inverse problem*, encountering several challenges that stem from noisy and incomplete data. The data usually encounters several different types of imperfections, measurement errors and reconstruction challenges [12].

1. First, to minimize radiation damage, the sample is usually exposed to a ***low electron dose***, which decreases the signal-to-noise ratio (SNR) of the micrographs. Specimen damage due to electron-specimen interaction is one of the major problems in Cryo-ET. Details regarding specimen damage due to ionization can be found in [13, 14]. The number of electron-specimen interactions need to be minimized such that the structural integrity of the specimen is preserved while collecting an entire tilt-series of images from the same specimen. This problem is addressed by reducing the dose on each individual tilt (projection) or by collecting less number of projections.

2. Second, ***shot noise*** can occur due to random arrival of electrons at a specific sensor element. Shot noise is Poisson-distributed and can be reduced using procedures of regularized refinement [15].

3. Third, 2D tilt images are usually collected at small angular differences. Ideally, tilts should be recorded to cover the entire 180° range of angles. However, due to the increased sample thickness at higher angles and the small space within the pole piece gap of the TEM objective lenses where the specimen is placed, the tilt angle is limited to ±70°. This leads to missing data commonly known as the ***missing wedge*** [16]. This turns the reconstruction inverse problem to be unstable without a unique solution rendering it ill-posed. The affect of having a missing wedge is shown in Fig. 1.7.

4. Fourth, ***specimen noise*** occurs due to rearrangement of the specimen during data recording. Although, the tilting process in modern microscopes is quite fine there can be mechanical errors.

5. Fifth, during data acquisition, the specimen is tilted, and at every tilt angle θ, its effective thickness increases according to $1/\cos\theta$ [17]. This increased thickness causes more inelastic scattering, which contributes to low contrast in the image.

6. Sixth, ***correlated noise*** can appear due to imperfections of the TEM detector. Normally, a gain reference is created to equalize the response from individual detector elements. However, errors in the gain reference give rise to noise that can be correlated with a region of the detector rather than with the specimen.

All these sources of imperfections, noise and missing data render the inverse problem severely ill-posed since the solution can not be unique and is unstable. A more detailed account of the unstable nature and non-uniqueness of inverse problems in general has been given in [18] and specifically for Cryo-ET reconstruction problem has been discussed in [12].

1.5 Tomographic Image Reconstruction

The word tomography originates from two Greek works 'tomos', meaning slice or part and 'graphein', meaning to write. Tomography is the art of reconstructing an object from its projections. Applications of tomographic reconstruction first came into consideration when it was realized that it was possible to non-invasively reconstruct the interior of a human body. There are a variety of tomographic modalities used in biomedicine, material science and geology. Most tomographic modalities are distinct on the basis of the imaging spectrum and the data acquisition geometry. All modalities are based on the premise that rays from a certain source passing through an object are altered, based on the alteration of the rays an image can be generated. The process is repeated by rotating the ray source or the object and the resulting images are processed to reconstruct a representation of the object. An interesting and detailed historical account of tomographic reconstruction can be found in [19]. In this section the review has been presented from the point of view of Cryo-ET that uses a parallel beam geometry.

1.5.1 Analytical Reconstruction Methods

Analytical reconstruction methods[1] are generally used to achieve relatively fast but primitive tomographic reconstructions. Results from such methods are often used for intuition and also serve as priors for more complex and computationally demanding iterative, statistical and model-based methods. Such methods are also important when computational time is critical and work well for applications where limited and low-dose data is not a major constraint. In certain applications a primitive but rapidly available reconstruction may be deemed sufficient. Certain aspects of these methods are still open research questions and entire books have been dedicated to analytical image reconstruction such as [20, 21].

There are several factors that limit the performance of analytical reconstruction methods. Such methods assume continuous measurements whereas sampling issues associated with analytical methods are dealt with in the discrete domain. Such methods also poorly handle measurement noise and noise removal is not incorporated in the problem formulation but is usually achieved as a post-processing step via filtering or other operations that also tend to remove useful information from the reconstruction. Removing streaking artifacts during post-processing is also fairly common as shown in [22, 23]. Although the limitations seem broad, such methods are commonly used in

[1] These methods are also refereed as *Fourier reconstruction methods* and *direct reconstruction methods.*

medical tomographic scanners that can collect sufficient and high contrast data using standard geometries and can avoid the limitations associated with such methods.

Radon Transform

The most fundamental component of analytical methods is the *radon transform* that was first presented in 1917 by Austrian mathematician *Johann Radon* in [24, 25]. Consider an object $f(x, y)$ as shown in Fig. 1.3 being imaged by parallel rays. A ray passes through the object at angle θ with the x-axis and has a distance of t_1 from the origin. The perpendicular line t_1 can be defined as,

$$t_1 = x\cos(\theta) + y\sin(\theta). \tag{1.2}$$

Fig. 1.3 represents only a single ray, *i.e.*, for $P_\theta(t_1)$, on collecting several rays one would end up with $P_\theta(t)$. The function $P_\theta(.)$ is called the *projection* of $f(x, y)$ at angle θ.

$$P_\theta(t) = \int_{x\cos(\theta)+y\sin(\theta)} f(x, y)\, ds \tag{1.3}$$

$$P_\theta(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\delta(x\cos(\theta) + y\sin(\theta) - t)\, dx\, dy, \tag{1.4}$$

where $\delta(.)$ donates the 1D Dirac impulse function. The radon transform can be used to model the *forward transform* and has been used in various forms in this thesis for simulating reconstruction problems.

Sinogram

A sinogram, S, is a matrix of all the projections collected from a specific object. In other words a sinogram is a discretized version of the two dimensional continuous Radon Transform. Fig. 1.4 shows the sinogram for a small box as well as a Shepp-Logan [26] phantom. The construction of the sinogram can be column-wise or row-wise, *i.e.*, each successive projection may be placed in a column or a row.

Fourier Slice Theorem

The Fourier Slice Theorem was first presented by Ron Bracewell in 1956 in [27] and is the foundation of analytical reconstruction methods[2]. If $f : \mathbb{R} \to \mathbb{C}$ is the continuous integrable function. The Fourier transform of $f(x, y)$ is defined as follows:

$$\hat{f}(u, v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)e^{-2\pi(ux+vy)i}\, dx\, dy \tag{1.5}$$

The inverse Fourier transform of $\hat{f}(u, v)$ is defined as follows:

[2] Although, the theorem first appears in the 1956 paper according to Jeff Fessler during a symposium held in July of 2004 at Stanford to celebrate the 75th birthday of Albert Macovski, Ron Bracewell mentioned that the theorem was commonly known among the radio astronomy community in the 1950s.

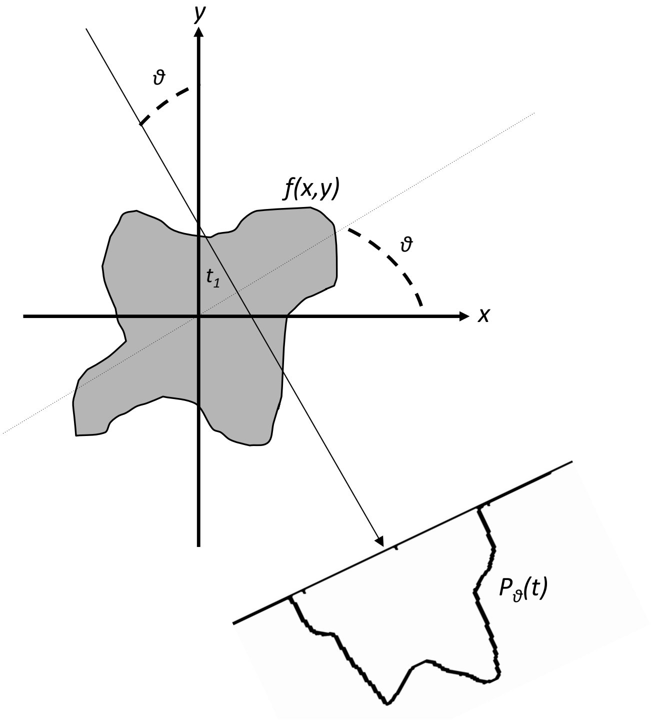


Figure 1.3: Figure showing graphical representation of the radon transform. A single ray passing through an object $f(x, y)$, when several rays pass through the $f(x, y)$ they form a projection of the object at an angle θ with the x-axis.

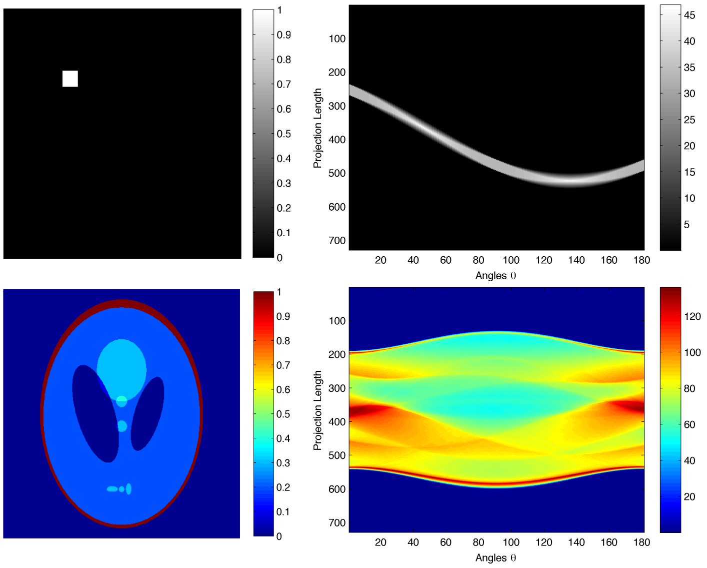


Figure 1.4: Figure showing graphical representation of the sinogram for a small box and the notorious Shepp-Logan phantom. In this case the sinogram is built column-wise, *i.e.*, each column represents a distinct 1D projection from the 2D object at a particular angle.

$$f(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \hat{f}(u,v)e^{2\pi(ux+vy)i}\,du\,dv \tag{1.6}$$

The Fourier slice theorem according to [21, 28] for reconstructing a 2D image from 1D projections can be stated as: the Fourier transform of a parallel projection of image $f(x,y)$ taken at an angle θ gives the vales of the two-dimensional transform $\hat{f}(u,v)$ along a line that is at an angle θ with the u-axis. This has been visually shown in Fig. 1.5. A primitive reconstruction can be obtained by taking successive Fourier transforms of projections and aligning them into a rotational grid and taking the inverse Fourier transform of the grid. From Fig. 1.5 it can be noted that higher frequency components, *i.e.*, points further away from the origin are less well known as compared to lower frequency components. These unknown points have to be approximated by interpolation. The interpolation problem has been addressed in detail in [21, 28]. Mathematically, the theorem can be defined as: *The Fourier transform of a parallel projection of a 2D object $f(x,y)$ taken at an angle θ $\hat{P}_\theta(\omega)$ gives the values of the 2D transform of $f(x,y)$, i.e., $\hat{f}(u,v)$ along a line that subtends and angle θ with the u-axis in the frequency domain [21],*

$$\hat{P}_\theta(\omega) = \hat{f}(\omega\cos(\theta), \omega\sin(\theta)). \tag{1.7}$$

Filtered Back Projection (FBP)

The filtered back projection algorithm and its variants can be constructed by rewriting the 2D Fourier transform presented above. It is evident from Fig. 1.5 that practically only a finite number of projections can be collected from a certain object. Aligning these projections in Fourier domain will thus always result in less information of higher frequency components as opposed to closely spaced lower frequency components. This essentially means that interpolation is always required which is much easier to do in the spatial rather than frequency domain. The FBP algorithm needs a change of coordinates from Cartesian to polar. Consider $u = \omega\cos(\theta)$ and $v = \omega\sin(\theta)$. The Jacobian $\mathcal{J}$ would be,

$$\mathcal{J} = \left|\frac{\partial(u,v)}{\partial(\omega,\theta)}\right| = \begin{vmatrix} \cos(\theta) & \sin(\theta) \\ -\omega\sin(\theta) & \omega\cos(\theta) \end{vmatrix} = \omega(\cos^2(\theta) + \sin^2(\theta)) = \omega. \tag{1.8}$$

Replacing the differentials $du\,dv$ to $\omega\,d\omega\,d\theta$, Equation 1.6 can be rewritten as

$$f(x,y) = \int_{0}^{2\pi}\int_{0}^{\infty} \hat{f}(\omega,\theta)\omega e^{2\pi(x\cos(\theta)+y\sin(\theta))i}\,d\omega\,d\theta \tag{1.9}$$

Simplifying further using the Fourier Slice Theorem and $t = x\cos(\theta) + y\sin(\theta)$ gives the following:

$$f(x,y) = \int_{0}^{\pi}\int_{-\infty}^{\infty} \hat{P}_\theta(\omega)\,|\omega|\,e^{2\pi\omega t i}\,d\omega\,d\theta \tag{1.10}$$

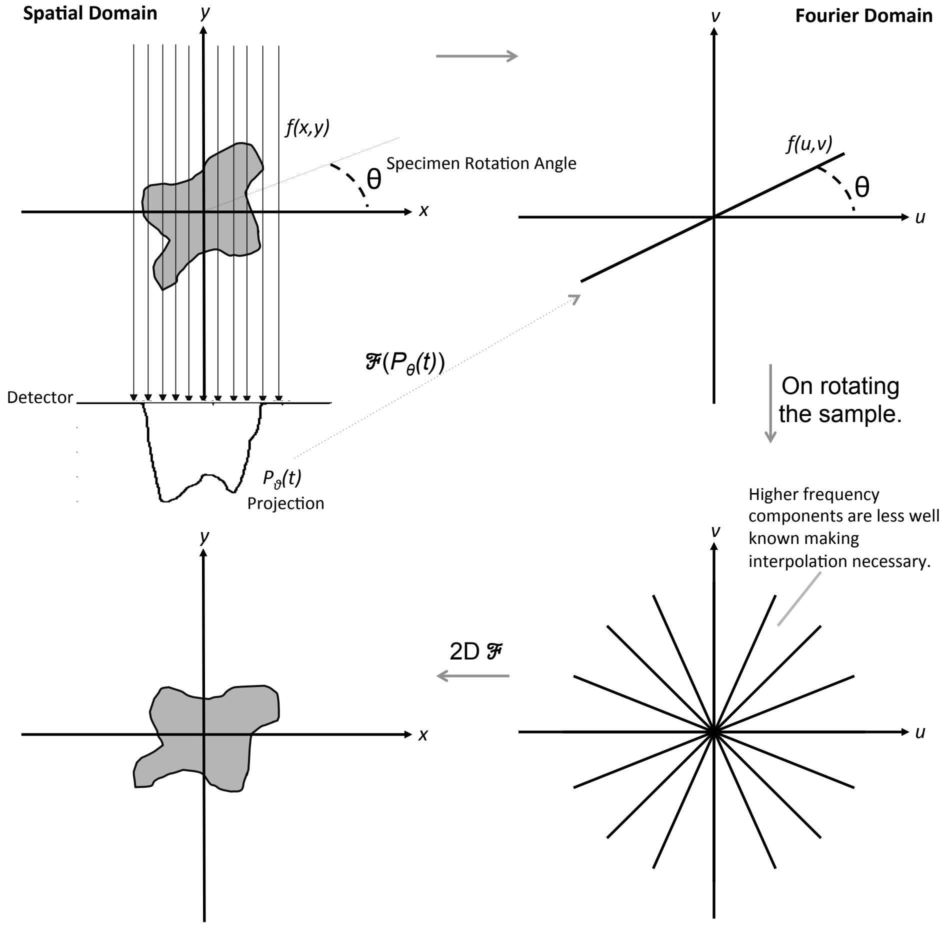


Figure 1.5: Figure showing the graphical representation of the projection slice theorem or Fourier slice theorem.

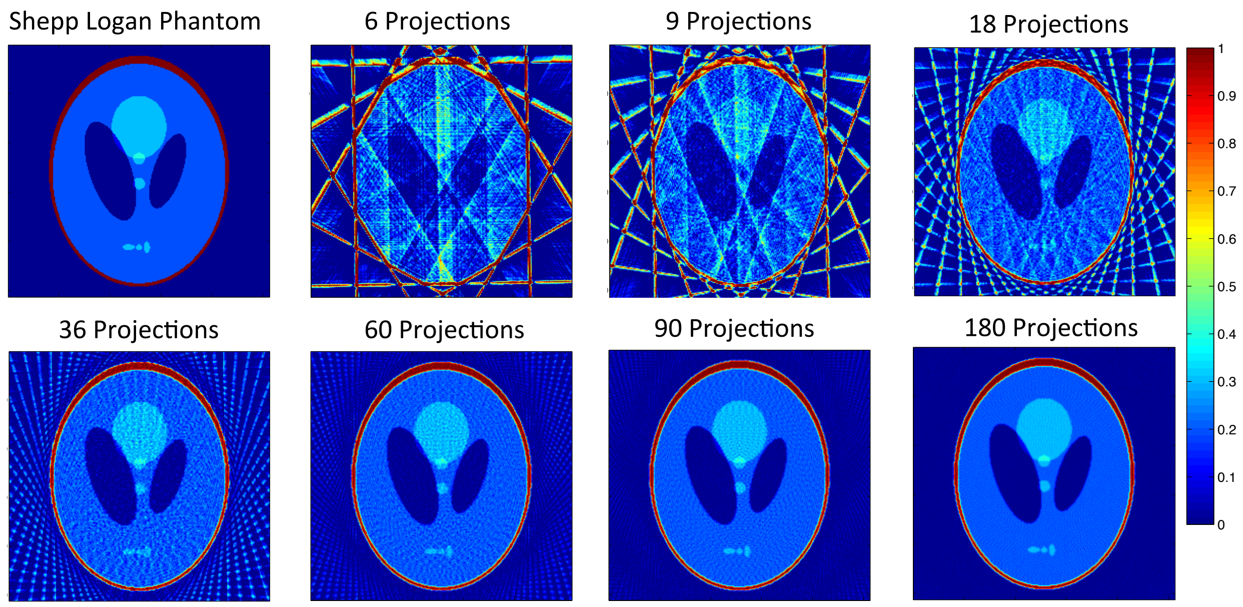


Figure 1.6: FBP reconstructions (Linearly interpolated, 'Ram-Lak' filter) with changing number of projections. The streaking artifacts show for lower number of projections show that there is not enough information to confine certain parts of the projections. The projections were simulated without any noise.

This equation explains fundamental FBP[3] The outermost integral represents the summations of the back projections over angles and the innermost integral is the back-projection of a single projection.

Fig. 1.6 shows results from reconstructing a modified Shepp-Logan phanton from 6,9,18,36,60,90 and 180 equally spaced projections[4]. Fig 1.7 shows FBP reconstructions with a missing wedge, this simulation is closer to a real Cryo-ET problem and the effects of the missing wedge are evident in both the spatial and frequency domain. A more detailed account of FBP and its variants can be found in [21].

1.5.2 Iterative Reconstruction Methods

As opposed to analytical methods which are dependent on step-by-step procedures as well as limited by geometrical constraints, iterative methods are often model-based, can incorporate prior information better and can also incorporate statistical noise models [7, 18, 29]. Development and enhancement of such methods is an active research area and this thesis partially focuses on the development, understanding and enhancement of such methods. Iterative methods can be classified as algebraic and variational methods. A variety of algebraic and variational methods are used in different chapters of

[3] In some literature this form of back-projection is also refereed to as weighted back projection (WBP), $|R|$-weighted back projection or according to the equation $|\omega|$-weighted back-projection.

[4] Although, reconstructions are generally performed on a circular grid for emulate perfect projections, in this thesis all reconstructions will be displayed on a rectangular grid, this is for the ease of visualization.

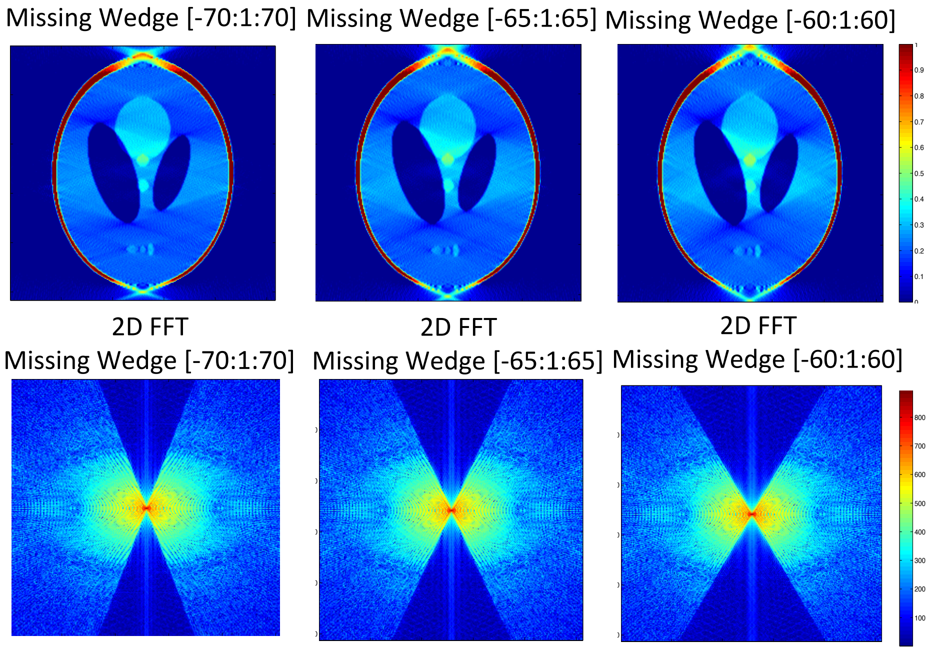


Figure 1.7: FBP reconstructions (Linearly interpolated, 'Ram-Lak' filter) showing the effect of the missing wedge in the reconstructed phantom as well as the frequency domain. The projections were simulated without any noise. The 2D FFT shown is with the zero frequency component shifted to the center. Higher frequencies are attenuated due to the Ram-Lak filter.

this thesis, fundamental methods have been explained as they are used in proceeding chapters.

Data Model: Discretization of the Cryo-ET Inverse Problem

In order to analyze the *ill-posed inverse problem* of reconstructing data from projections (tilt-series), it is fundamental to have an accurate formulation of the *forward operator*, solving the problem of modeling the process of image formation in the absence of noise and measurement errors [11]. The collected projection images can be denoted by:

$$P_\theta : 0 \leq x < x_{max}, 0 \leq y < y_{max}, -ma^\circ \leq \theta \leq +ma^\circ \tag{1.11}$$

Where x_{max} and y_{max} is the maximum size of the projections and $\pm ma^\circ$ is the maximum possible tilt angle, these parameters are determined by the microscope hardware [30]. The forward model for Cryo-ET has been extensively discussed in [11] where it has been shown that it may be assumed that collected tilt series images are equivalent to the forward projection, *i.e.*, Radon Transform of the specimen. A sinogram (S) represents the tilt series of raw data, *i.e.*, a matrix where each column represents a projection at a different angle. Mathematically, a discretized and simplified version of the noise-free forward operator can be described in terms of a linear system where projections b from are collected from specimen x, given a matrix representation of the imaging device A.

$$Ax = b \qquad A \in \mathbb{R}^{m \times n} \;\; x \in \mathbb{R}^n \;\; b \in \mathbb{R}^m \tag{1.12}$$

The vector b is the vectorized form of S which is constructed from projections P_θ. Each row of A corresponds to a single ray passing through the density being imaged. Since each ray only passes through a certain number of voxels, matrix A is usually sparse. This sparsity can be used to a computational advantage since A can be stored and used in a sparse way, utilizing less memory.

The imaging model presented above represents a set of linear equations such that there is one equation for each ray passing through the object. In an ideal, noise-free case, these linear equations would be consistent. However, in practice this is never the case, since the right hand side of the linear system is usually marred by noise $b = b^* + e$, where b^* is ideal data and e represents perturbation[5] or data errors. Throughout this thesis we assume that all the forms of measurement errors and missing data limitations are reflected in b.

Algebraic Reconstruction Methods

Algebraic Reconstruction Methods (ARMs)[6] are a class of methods which solve Equation 1.12 as a set of linear equations using standard equation solving methods such as Kaczmarz, Landwabber, Cimmino etc. Kaczmarz method in particular was reinvented as the Algebric Reconstruction Technique (ART) specifically for tomographic applications. Many of these methods are stopped at semi-convergence for stability, hence the iteration index acts as a regularization parameter. A more detailed description of ARMs has been given in Chapter 4 where they are used for extended field-based reconstructions.

Variational Reconstruction Methods

Variational methods re-cast the reconstruction problem as an optimization problem that involves one or multiple forms of regularization. As mentioned earlier the right-hand-side of $Ax = b$ is corrupted by noise and measurement errors mentioned in section 1.4. Although algebraic methods such as ART and SIRT do possess some regularization capabilities most of them are derived from semi-convergence [18, 31]. To handle arbitrarily large perturbations on the right-hand-side of Equation 1.12 we need regularization that can compute more sensitive approximations. The typical approach is to formulate the optimization problem as a minimization of the linear combination of a data consistency error term and a regularization penalty as,

$$\operatorname*{argmin}_x D(A(x), b) + \lambda G(x). \tag{1.13}$$

[5] The word *perturbation* used in this thesis refers to secondary influence on an imaging system that causes deviation.

[6] ARMs or Row Action Methods (RAMs) and methods associated with them have been ambiguously defined in various literature and are often a source of confusion. The term ARM is often confused with ART, ARMs represent a class of methods while ART is simply Kaczmarz Method used in a tomographic setting.

In the above optimization problem $x \to G(x)$ is the regularization term with λ as the regularization parameter. $D(A(x), b)$ is the data discrepancy functional or the data consistency error term. A more detailed account of fundamentals of variational reconstruction methods can be found in [18].

1.6 High-Performance and Reconfigurable Computing

All methods discussed in the previous sections are computationally demanding specifically for large size reconstructions. Speeding-up analytical reconstruction methods and decreasing the convergence time for iterative methods are active research areas. Analytical methods can provide a primitive reconstruction that can then be used as a starting point (prior) for more complex iterative methods and help speed up convergence. Although, analytical methods such as FBP perform relatively faster than iterative methods they become increasingly complex for large image sizes since their complexity scales according to $\mathcal{O}(n^3)$)).

Although, the number of transistors in a dense integrated circuit doubles every two years [32] our hunger for computational speed, throughput and performance increases proportionally. For every general purpose processing system there is a problem, which is too large for it to handle, this is where application specific hardware comes in and provides solutions, which are robust and are based on handling several tasks in a single computation cycle. Enhancing performance by customizing hardware comes with a standard trade-off of losing flexibility. Reconfigurable computing (RC) combines the high performance of a hardware system with the flexibility of software and bridges the gap between general-purpose processors (GPPs) and Application Specific Integrated Circuits (ASIC) [33, 34]. Estrin and Bussell introduced reconfigurable computing in 1963 [35] but it has not been studied in depth because commercial reconfigurable hardware only became available in the 1990s. International Technology Road-map for Semiconductors has predicted that the minimum feature size for modern chips will reach as low as 5nm and on chip clock frequency of about 50 GHz by 2020. The flexibility and reprogramming ability of general-purpose computers served as the essence of the computing revolution. Very large-scale integration (VLSI) based technologies enabled us to integrate entire processors into single chips. The fact that a single processor could perform a variety of operations without hardware modification was revolutionary but it came at the expense of several trade-offs including speed and throughput [35]. Nowadays, clock frequencies have reached up to 4GHz, which means GPPs, are capable of performing high-speed computations in a single clock cycle. Reconfigurable computing can bridge the gap between hardware and software; so as to obtain highly efficient performance than pure software implemented microprocessors, while being more flexible than ASIC hardware.

Trade-off between high performance and flexibility has always been a problem in modern digital design, the choice between a completely inflexible but efficiently performing ASIC on one end and a highly flexible GPP with inferior performance. Although FPGAs run at much lower clock speeds they have the tendency to have higher computational speeds because they require much less cycles to execute an operation as compared to GPPs. RC is a unique computing model, which allows post-manufacturing reconfiguration. Nowadays Graphical Processing Units (GPUs) are becoming more and more common to achieve high throughput on large scale computing tasks. However, they require more energy and the transition between a GPU implementation and an ASIC implementation would be relatively more tedious as opposed to moving from an FPGA design to ASIC implementations.

Application Specific Integrated Circuits (ASICs) are hardwired systems designed to perform a certain task, although they rank at the highest level on the performance scale they have very little to negligible flexibility. Although it may seem to be an ultimate solution for achieving high throughput design, ASICs entail several issues from design complexity to having high degree of design failure and testing problems. Reconfigurable Computing acts as the most optimal solution since it does not require a chip being manufactured every time a certain design has to be tested while at the same time it provides a significant degree of higher performance as compared to GPPs. The most modern form of reconfigurable computing are Field Programmable Gate Arrays (FPGAs) that have evolved from Programmable Logic Devices (PLDs). Despite having clock cycle speeds several times less than GPPs, FPGAs can achieve much higher speeds because they are able to reduce the overall number of cycles required to execute a certain task. This will further be discussed in more detail in Chapter 5.

1.7 Contributions

Figure 1.8 reflects the contributions of this thesis. The basic aim of the thesis is to explore algorithmic methods for improving tomographic reconstructions and the development of reconfigurable computing-based architectures to speed-up and accelerate such methods. Major Contributions Include:

1. Studying Graph-based non-local methods and their feasibility for improving tomographic reconstructions. Using the ability of graph-based non-local Total Variation to establish a connection between different regions of an image which are spatially far from each other to refine the sinogram by exploiting the structure within the sinogram *(Chapter 2)*.

2. Development of adaptive graph-based total variation (AGTV) for tomographic reconstructions Similar to state-of-the-art Non-local TV (NLTV) the proposed method goes beyond spatial similarity between different regions of an image being reconstructed by establishing a connection between similar regions in the image regardless of spatial distance. However, it involves updating the graph prior during every iteration making the connection between similar regions stronger. Moreover, it promotes sparsity in the wavelet and graph gradient domains. Since TV is a special case of graph TV the proposed method can also be seen as a generalization of Sparsity Exploiting Reconstruction (SER) and TV methods *(Chapter 3)*.

3. Studying the affect of extending field (EF)-based noise removal from tomographic reconstructions. The study demonstrates that extending the reconstruction space,

which increases the dimensionality of the linear system being solved during iterative tomographic reconstruction, facilitates the separation of signal and noise. A considerable amount of the noise associated with collected projection data arises independently from the geometric constraint of image formation, whereas the solution to the reconstruction problem must satisfy such geometric constraints. Increasing the dimensionality thereby allows for a redistribution of such noise within the extended reconstruction space, while the geometrically constrained approximate solution stays in an effectively lower dimensional subspace. Employing various tomographic reconstruction methods with regularization capability we performed extensive simulation and testing and we observed that enhanced dimensionality significantly improves the accuracy of the reconstruction. Although the proposed method is used in the context of Cryo-ET, the method is general and can be extended to a variety of other tomographic modalities *(Chapter 4)*.

4. The two-dimensional Discrete Fourier Transform is a fundamental component of a variety of methods used for Cryo-ET. All methods developed in contributions 1-3 assume Filtered Back Projection (FBP) as a starting point of prior. Although FBP is a computationally efficient operation when compared to iterative reconstruction methods it can be computationally demanding for large size reconstructions. One of the major bottlenecks for computing the FBP is the 2D FFT. 2D FFTs also have a variety of other applications for Cryo-ET e.g. in subtomogram averaging etc. We develop an FPGA based 2D FFT with simultaneous edge artifact removal. The contributions in the reconfigurable computing domain are two-fold:

   - Since the 2D FFT is generally implemented on FPGAs using a row-column decomposition (RCD)-based approach intermediate storage and access from the external DRAM becomes a problem for large datasets. We propose a tile-hopping memory mapping scheme which minimizes strided DRAM access and can increase the overall bandwidth of data exchange from external memory *(Chapter 5)*.
   - Since we incorporate simultaneous edge artifact removal into our 2D FFT implementation this results in added complexity. We propose an optimized periodic plus smooth decomposition-based approach which reduces 1D FFT invocations and DRAM access *(Chapter 5)*.

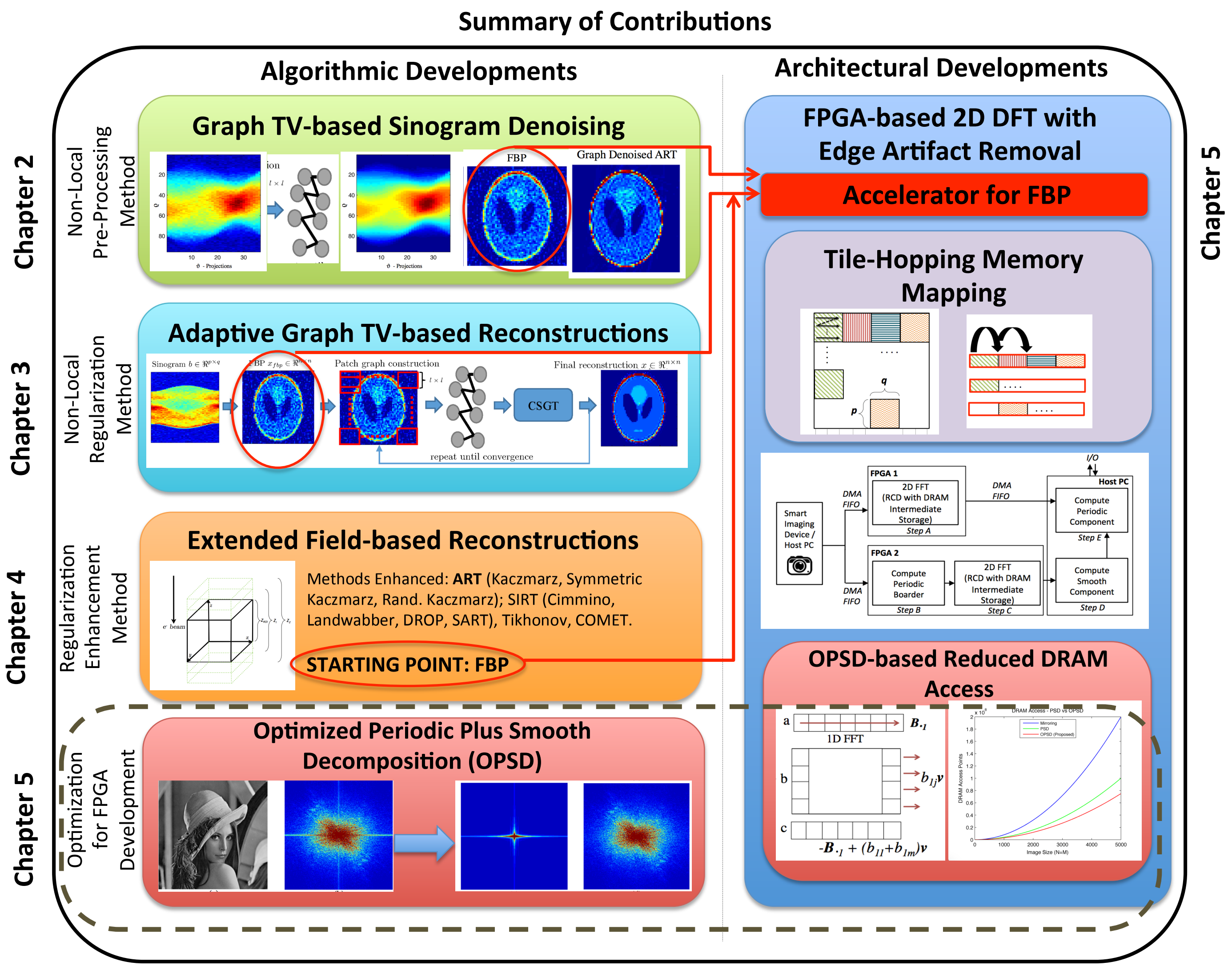

Summary of Contributions
Algorithmic Developments
Architectural Developments
Chapter 2
Non-Local Pre-Processing Method
Graph TV-based Sinogram Denoising
FBP
Graph Denoised ART
θ - Projections
Chapter 3
Non-Local Regularization Method
Adaptive Graph TV-based Reconstructions
Sinogram $b \in \Re^{p\times q}$
FBP $x_{fbp} \in \Re^{n\times n}$
Patch graph construction
$l \times l$
CSGT
Final reconstruction $x \in \Re^{n\times n}$
repeat until convergence
Chapter 4
Regularization Enhancement Method
Extended Field-based Reconstructions
Methods Enhanced: ART (Kaczmarz, Symmetric Kaczmarz, Rand. Kaczmarz); SIRT (Cimmino, Landwabber, DROP, SART), Tikhonov, COMET.
STARTING POINT: FBP
Chapter 5
Optimization for FPGA Development
Optimized Periodic Plus Smooth Decomposition (OPSD)
FPGA-based 2D DFT with Edge Artifact Removal
Accelerator for FBP
Tile-Hopping Memory Mapping
p
q
Smart Imaging Device / Host PC
DMA FIFO
FPGA 1
2D FFT (RCD with DRAM Intermediate Storage) Step A
FPGA 2
Compute Periodic Boarder Step B
2D FFT (RCD with DRAM Intermediate Storage) Step C
Host PC
I/O
Compute Periodic Component Step E
Compute Smooth Component Step D
OPSD-based Reduced DRAM Access
a
1D FFT
$B_{\cdot 1}$
b
$b_{1j}v$
c
$-B_{\cdot 1} + (b_{11}+b_{1m})v$
DRAM Access - PSD vs OPSD
Mirroring
PSD
OPSD (Proposed)
DRAM Access Points
Image Size (N=M)
Chapter 5

Figure 1.8: A graphical representation of algorithmic and architectural contributions.

Chapter 2

Graph-based Non-Local Sinogram Denoising

2.1 Chapter Outline

Traditionally, denoising the raw projection data acquired from various tomographic modalities as a pre-processing step has been discouraged due to high degree of manipulation associated with standard filtering and other denoising methods. Observing the raw data in terms of the sinogram shows that it has structure and high redundancy since each projection is collected from the same object. In this chapter, a novel algorithm for denoising the sinogram is presented inspired by the modern field of signal processing on graphs. This method has been published in [36]. Graph-based non-local methods often perform better than standard filtering operations since they can exploit the signal structure by establishing a connection between regions of the data that are spatially far from each other. Non-local image processing has the ability to prevent over-smoothing and can preserve texture details better than local methods. This makes the sinogram an ideal candidate for graph-based non-local denoising given its high degree of redundancy and the fact that any denoising method employed should preserve high-frequency details. The proposed method was tested with a variety of phantoms and different reconstruction methods. The presented numerical study shows that the proposed algorithm improves the performance of analytical filtered back-projection (FBP) and iterative methods ART (Kaczmarz) and SIRT (Cimmino). It was observed that a graph denoised sinogram always minimizes the error measure, improves the accuracy of the solution with minimal resolution loss as compared to regular analytical and iterative methods that attempt to denoise the data as an integral component of the reconstruction method.

2.2 Sinogram Denoising

As extensively discussed in Chapter 1 low-dose, limited, and incomplete data are common problems in a variety of tomographic reconstruction paradigms including Cryo-ET. Low dose computerized tomography (CT) and Cryo-ET reconstructions are *ill-posed inverse problems* that encounter significant amounts of noise [16, 37]. During data collection, individual projections are usually marred by noise due to low illumination, which stems from low electron or x-ray doses.

Current efforts to reduce the dose for Cryo-ET reconstructions can be divided into three categories [38][39]:

1. Pre-processing-based methods, that tend to improve the raw data (sinogram), followed by standard analytical or iterative reconstruction methods [40, 41].

2. Denoising of reconstructed tomograms in the image domain. This approach is usually accomplished by filtering or extensive averaging. Post-processing of reconstructed densities to enhance features, decrease limited angle streaking artifacts and removing speckle (*i.e.,* areas with inconsistent intensity) is particularly common [23, 42].

3. Iterative image reconstruction and statistical methods [43]. Such methods are often based on variational regularization and recast the reconstruction problem as an optimization problem. This optimization problem is then solved by moving back and forth between the sinogram and reconstruction domain so that the final result is faithful to the raw data while satisfying a regularization penalty term to constrain the noise.

A diagram representing various methods to achieve better tomographic reconstructions has been shown in Fig. 2.1. Traditionally, the idea of manipulating raw data before reconstruction has been treated with caution since there is a possibility of removing high frequency signal information from raw data while removing noise. However, recent developments in image denoising methodology have changed this. A review of recent image denoising algorithms has been presented in [44]. There has been a growing interest in denoising the sinogram, specifically for CT reconstructions [45, 46]. These algorithms vary from simple adaptive filtering and shift-invariant low-pass filters to computationally complex Bayesian methods [46, 47]. Other approaches involve Fourier and wavelet transform-based, multi-resolution methods [21] and denoising projection data in Radon space [48]. Since the variance of Poisson noise is dependent on the amplitude of the signal, adaptive filters seem to be a good choice for denoising the sinogram and have been studied in [42, 49]. Methods that denoise the projection data in Radon space (after logarithmic transformation) are of particular interest because Poisson noise can be treated as Gaussian making noise refinement easier. This has been theoretically studied in [48, 50]. Over the past few years there has been a growing interest in graph-based non-local methods for image processing applications. Such methods are of particular interest because of their ability to establish connections between regions of the image that are spatially far from each other. This helps in processing similar regions together rather than the traditional approach of processing proximal pixels. In this study a non-local graph-based sinogram denoising approach is presented and analyzed.

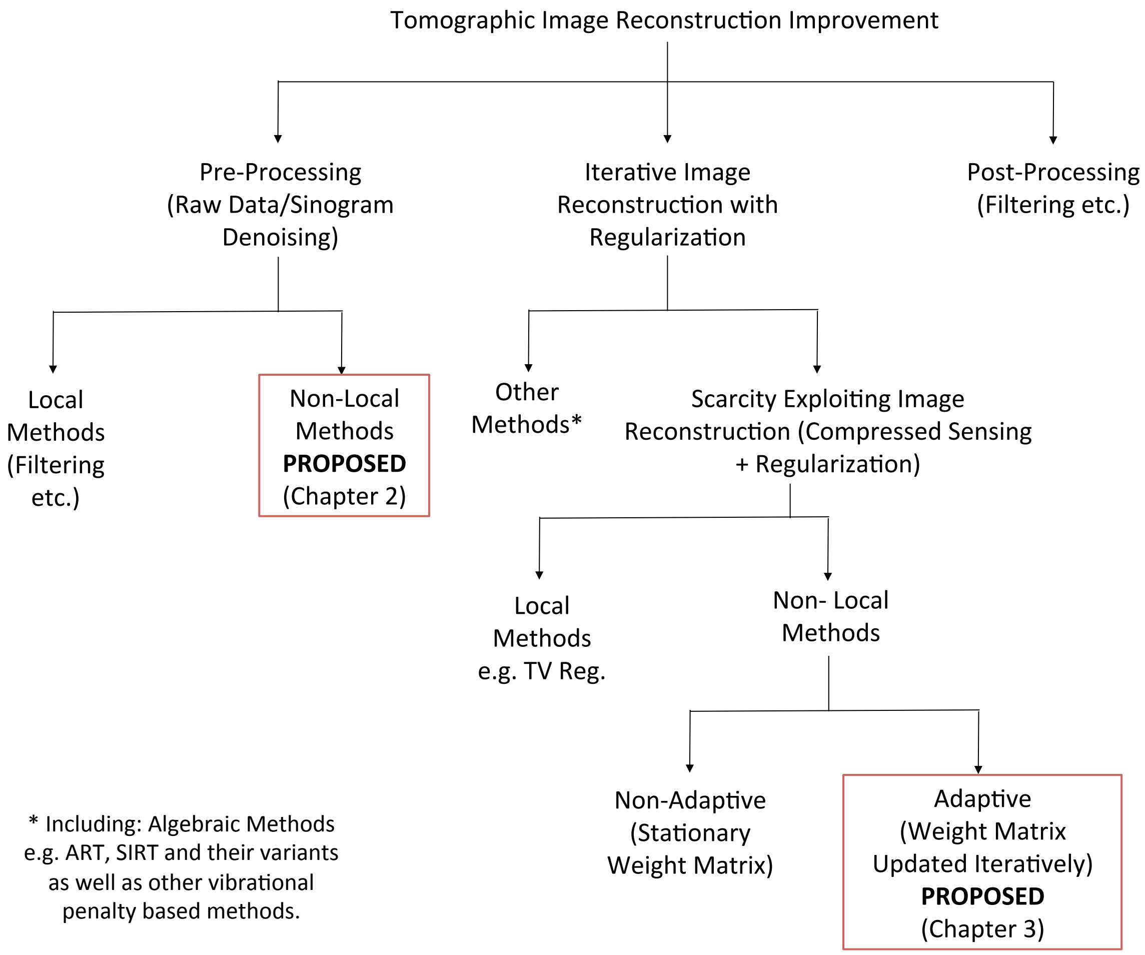

Tomographic Image Reconstruction Improvement
Pre-Processing (Raw Data/Sinogram Denoising)
Iterative Image Reconstruction with Regularization
Post-Processing (Filtering etc.)
Local Methods (Filtering etc.)
Non-Local Methods **PROPOSED** (Chapter 2)
Other Methods*
Scarcity Exploiting Image Reconstruction (Compressed Sensing + Regularization)
Local Methods e.g. TV Reg.
Non- Local Methods
Non-Adaptive (Stationary Weight Matrix)
Adaptive (Weight Matrix Updated Iteratively) **PROPOSED** (Chapter 3)
* Including: Algebraic Methods e.g. ART, SIRT and their variants as well as other vibrational penalty based methods.

Figure 2.1: Various tomographic reconstruction improvement methods summarized. Most methods are based on local image processing paradigms whereby only proximal pixels or voxels are taken into account while processing the data. This chapter focuses on a non-local pre-processing method which is proceeded by a non-local sparsity exploiting regularization-based method in Chapter 3.

2.3 Introduction to Graphs

Graphs describe geometric structures of data domains in a plethora of applications such as energy systems, transportation, sensor networks, human neuronal networks. The edge weights of a graph reflect the similarity between the two vertices the edge connects. The edge weights depend on the problem under consideration. In image processing there has been an increasing interest in graph-based non-local or semi-local processing specifically for filtering [51, 52]. Such approaches are able to connect pixels of an image that are not only in the regional proximity but spatially far for each other. Thus non-local methods have the ability to better preserve edges and textures in an image. A more detailed review regarding graph-based methods can be found in [53, 54].

Signal Processing on Graphs Preliminaries

Mathematically, a graph is represented as a tupple $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathcal{W}\}$ where $\mathcal{V}$ is a set of vertices, $\mathcal{E}$ a set of edges, and $\mathcal{W} : \mathcal{V} \times \mathcal{V} \to \mathbb{R}_+$ a weight function. We assume that the vertices are indexed from $1, \ldots, |\mathcal{V}|$. The weight matrix W is assumed to be non-negative, symmetric, and with a zero diagonal. Each entry of the weight matrix $W \in \mathbb{R}_+^{|\mathcal{V}|\times|\mathcal{V}|}$ corresponds to the weight of the edge connecting the corresponding vertices: $W_{i,j} = \mathcal{W}(v_i, v_j)$ and if there is no edge between two vertices, the weight is set to 0. A node v_i connected to v_j is denoted by $i \leftrightarrow j$. For a vertex $v_i \in \mathcal{V}$, the degree $d(i)$ is defined as the sum of the weights of incident edges: $d(i) = \sum_{j\leftrightarrow i} W_{i,j}$. We define a graph signal as a function $s : \mathcal{V} \to \mathbb{R}$ which assigns a value to each vertex in the graph. It is convenient to consider a signal s as a vector of size $|\mathcal{V}|$ with the i^{th} component representing the signal value at the i^{th} vertex.

We define two types of signals on graphs, the one which resides on the vertices and the other one on edges. For a signal s residing on the vertices of graph $\mathcal{G}$, the gradient $\nabla_{\mathcal{G}} : \mathbb{R}^{|\mathcal{V}|} \to \mathbb{R}^{|\mathcal{E}|}$ is defined as

$$\nabla_{\mathcal{G}} s(i,j) = \sqrt{W(i,j)} \left(\frac{s(j)}{\sqrt{d(j)}} - \frac{s(i)}{\sqrt{d(i)}} \right),$$

where we consider only the pair $\{i,j\}$ when $i \leftrightarrow j$. For a signal f residing on the graph edges, the adjoint of the gradient $\nabla_{\mathcal{G}}^* : \mathbb{R}^{|\mathcal{E}|} \to \mathbb{R}^{|\mathcal{V}|}$, called divergence can be written as

$$\nabla_{\mathcal{G}}^* f(i) \quad = \quad \sum_{i\leftrightarrow j} \sqrt{W(i,j)} \left(\frac{1}{\sqrt{d(i)}} f(i,i) - \frac{1}{\sqrt{d(j)}} f(i,j) \right).$$

The Laplacian corresponds to the second order derivative and its definition arises from $Ls := \nabla_{\mathcal{G}}^* \nabla_{\mathcal{G}} s$. Let D be the diagonal degree matrix with diagonal entries $D_{ii} = d(i)$, then another interpretation of the Laplacian is the difference of the weight matrix W from the degree matrix D, thus $L = D - W$, which is referred to as combinatorial Laplacian.

A graph $\mathcal{G}$ can be constructed from a dataset in several ways. For example, for a signal $Y \in \Re^{p\times n}$, the vertices v_i of the graph $\mathcal{G}$ can correspond to the data samples $y_i \in \Re^p$. For the purpose of this work we are interested in a standard $\mathcal{K}$-nearest neighbors strategy for the graph construction. The first step of this strategy consists of searching the closest neighbors for all the samples using Euclidean distances. Thus, each y_i is connected to its $\mathcal{K}$ nearest neighbors (nearest in term of Euclidean distance not proximal distance) y_j, resulting in $|\mathcal{E}| \approx \mathcal{K}n$ number of connections. The second step consists of computing the graph weight matrix W using one of the several commonly used strategies. The most commonly used scheme is based on the Gaussian kernel:

$$W_{ij} = \begin{cases} \exp\Big(- \frac{\|(y_i - y_j)\|_2^2}{\sigma^2} \Big) & \text{if } y_j \text{ is connected to } y_i \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

Other common weighting schemes include the binary weighting kernel and correlation based weighting. Binary weighting can not give importance to each patch, correlation-based schemes are generally used for temporal signals (e.g. EEG signals). Gaussian-based weighting is the most common weighting scheme for imaging applications.

It should be noted here that the nearest neighbor parameter, $\mathcal{K}$, has to be selected carefully. If $\mathcal{K}$ is too large the graph would be fully connected and may lead to over-smoothing. If $\mathcal{K}$ is too low less similarity between different regions of the image will be considered. Unfortunately, there is no hard and fast rule to tune $\mathcal{K}$ except for hit and trial.

How can the notion of graphs be used in signal processing? A primary example is the denoising operation, where one can solve an inverse problem by regularizing with a *graph tikhonov operator.* For a noisy signal $y \in \mathbb{R}^n$ and a graph Laplacian L constructed between the entries of y, the graph tikhonov regularization on the optimization variable x (the clean signal) is defined as [53]:

$$x^\top L x = \frac{1}{2} \sum_i \sum_j W_{ij}(x_i - x_j)^2.$$

Clearly, the strongly connected samples in y_i, y_j have a higher W_{ij}, therefore, the resulting x_i, x_j resemble each other more. This smoothness prior can be used to denoise a signal y with an irregular structure by solving the optimization problem of the following form:

$$\min_x \|y - x\|_2^2 + \gamma x^\top L x$$

GTV is another graph based regularizer that has been frequently used in many applications. We discuss this regularizer in detail in the next section.

Why is the concept of graphs interesting for image processing applications? Images generally contain significant self-similarities on a pixel level. Most image processing paradigms focus on processing adjacent pixels of an image for refinement, filtering, information extraction and analysis. However, pixel-level similarities transcend spatial proximities, regions of a specific image may be similar to pixels that are spatially far from each other. Non-Local methods[1] exploit this similarity by processing similar

[1] The major difference between non-local methods and graph-based methods for image processing is the use of nearest neighbor search conditions. Due to the relative newness of the field the term non-local methods is often used synonymous to patch-based methods or graph-based methods for image

patches of an image together.

The non-local means (NLM) filter introduced by *Buades et.al.* in [51, 55] was the first attempt to denoise images non-locally. NLM takes advantage of similar patches in an image which are non-local to each other, *i.e.*, spatially far from each other. The similarity between the patches is given a weight, thus the similarity can be represented as a weight function inducing a weighted averaging filter. The well known bilateral image filter can be considered to be a special case of the NLM filter. NLM simply uses self similarity to reduce the noise by averaging the near-by (but not essentially spatially proximal) similar weighted overlapping patches.

Most non-local algorithms are designed for general denoising and are not applicable to inverse problems. In this chapter we focus on denoising the raw data (sinogram) using non-local total variation, this work is proceeded by development of an adaptive graph-based reconstruction algorithm in Chapter 3. The proposed method can exploit the hidden structure of the sinogram due to high redundancy of information in projections collected from the same sample.

2.4 Graph-based Sinogram Denoising

Graph-based denoising, tends to perform better than simple filtering operations due to its inherent capacity to exploit signal structure [53]. The proposed denoising method is general in the sense that it can be applied to raw data independent of the type of data, tomographic modality under consideration or the reconstruction method being employed.

Preliminaries

Let $S \in \Re^{p \times q}$ be the sinogram corresponding to the projections of the sample $x \in \Re^{n \times n}$ being imaged, where p is the number of rays passing through x and q is the number of angular variations at which x has been imaged. Let $b \in \Re^{pq}$ be the vectorized measurements or projections and $A \in \Re^{pq \times n^2}$ be the sparse projection operator. Then, the goal of a typical tomographic reconstruction method is to recover the sample x from the projections b. Of course, one needs to solve a highly under-determined inverse problem for this type of reconstruction, which can be even more challenging if the projections b are noisy. To circumvent the problem of noisy projections, a proposed two-step methodology for the reconstruction is given below:

1. *Denoise the sinogram, S, using graph total variation regularization.*
2. *Reconstruct the sample, x, from denoised projections using standard reconstruction methods.*

We motivate the denoising method here via Fig. 2.2(a,b,d,e), which shows sinograms corresponding to the Modified Shepp-Logan and Smooth phantoms. It can

processing. However, this can be a source of confusion since there is no universally accepted definition for how the weight matrix is constructed. For the purposes of this thesis we define graph-based methods as mentioned in section 2.3

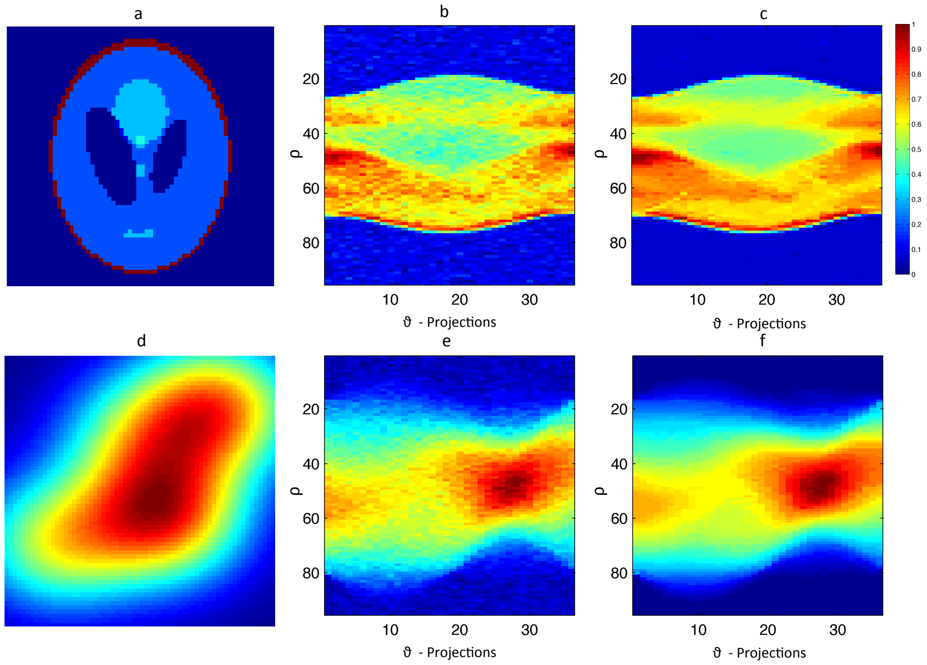


Figure 2.2: a) The Shepp Logan Phantom. b) Noisy sinogram of the Shepp Logan phantom (relative noise 0.08) c) Graph Denoised Sinogram. d) Smooth Phantom. e) Noisy sinogram of the Smooth phantom (relative noise 0.05) f) Graph Denoised Sinogram. Clearly, the projections have a smooth structure, which when exploited, denoises these sinograms significantly.

be observed that the sinograms have a piecewise smooth structure. If a sinogram is treated as an image, it can be said that some patches of the sinogram are similar to other patches for a given phantom. This structure can be exploited in the form of a pairwise similarity graph constructed between the patches of the sinogram and then used to denoise it via *graph regularization*.

2.4.1 Graph Construction for Sinogram Denoising

In this case the graph $\mathcal{G}$ is a patch graph, *i.e*, a graph between the patches of sinogram S and it is built using a three-step strategy. In the first step the sinogram $S \in \Re^{p\times q}$ is divided into pq overlapping patches. Let s_i be the patch of size $l \times l$ centered at the i^{th} pixel of S and assume that all patches are vectorized, *i.e*, $s_i \in \Re^{l^2}$. In the second step the search for the closest neighbors for all vectorized patches is performed using the Euclidean distance metric. Each s_i is connected to its $\mathcal{K}$ nearest neighbors s_j, resulting in $|\mathcal{E}|$ number of connections. In the third step the graph weight matrix W is computed as defined in 2.1.

The parameter σ, used while calculating the weights, can be set experimentally as the average distance of the connected samples. This procedure has a complexity of $\mathcal{O}(pqe)$ where e is the number of edges which scale linearly with the size of the image (sinogram in this case). Although the complexity seems to be a major issue it can be reduced by using approximations. This will be discussed in more detail in Chapter 3.

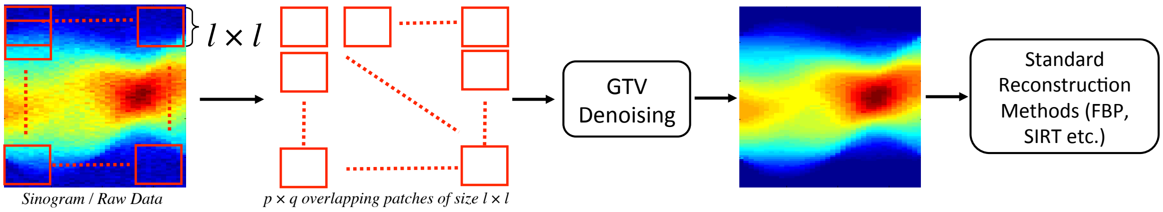

$l \times l$
Sinogram / Raw Data
$p \times q$ overlapping patches of size $l \times l$
GTV Denoising
Standard Reconstruction Methods (FBP, SIRT etc.)

Figure 2.3: The overall process of Graph-based Sinogram denoising showing graph construction on the sinogram.

2.4.2 Graph Total Variation Denoising

Once the graph $\mathcal{G}$ has been constructed, we solve the following optimization problem to denoise b (*i.e.* the vectorized form of the sinogram).

$$\min_z \|z - b\|_2^2 + \gamma \|\nabla_{\mathcal{G}} z\|_1, \tag{2.2}$$

where z is the clean vectorized sinogram. The problem can also be written as,

$$\min_z \|z - b\|_2^2 + \gamma \sum_i \sum_j \sqrt{W_{i,j}} |z_i - z_j|, \tag{2.3}$$

where $\nabla_{\mathcal{G}} z$ denotes the graph total variation of z. The parameter γ controls the amount of smoothing on z. Higher levels of noise require higher γ but at the expense of losing signal information. The choice of γ is heavily dependent on the type of data and the amount of noise in the sinogram. This parameter generally has to be tuned for a certain type of signal. The optimization problem is referred to as Graph Total Variation used here as a denoising operator, and minimizes the sum of the gradients at the nodes of the graph. This setup is similar to standard non-local TV regularization first presented by *Rubin, Osher and Fatemi* in the so called ROF model in [56]. However, in this case it is applied as a TV filter to denoise the sinogram rather than solving an actual inverse problem and also involves graph-based non-local processing. A more detailed account of Graph TV has been explained in the proceeding chapter (Section 3.4.3) where it is used as an additional sparcifying transform for solving an inverse problem.

Stated more simply, as ℓ_1 norm promotes sparsity, problem (2.2) tends to smooth the projections b, such that the new projections z have sparse graph gradients. Intuitively, this makes sense, as the sinograms (Fig. 2.2) have a piecewise smooth structure. *Thus, all the strongly connected patches in S will have a similar structure in z (zero graph gradients) whereas the weakly connected patches in S will have different structure in z (non-zero graph gradients).* It should be noted here that the weights in equation 2.3 are from the patches while the ℓ_1 distance is between pairs of pixels. Weights are assigned on pairs of patches and not between pairs of pixels to avoid assigning weights to noise.

The solution to problem (2.2) is given by the Algorithm 1. The algorithm requires

Algorithm 1 Graph Total Variation Denoising

```
INPUT: $u_0 = 0$, $\epsilon > 0$, OUTPUT: $z_j$
for $j = 0, \ldots J-1$ do
    $z_j = b - \nabla^*_{\mathcal{G}}(u_j)$
    $r_j = u_j\tau + \nabla_{\mathcal{G}}(z_j)$
    $s_j = \max(r_j - \gamma\tau, 0)$
    $u_{j+1} = \frac{1}{\tau}(r_j - s_j)$
    $F_{j+1} = \|b - z_j\|_2^2 + \gamma\|u_{j+1}\|_1$
    if $\frac{\|F_{j+1} - F_j\|_F^2}{\|F_j\|_F^2} < \epsilon$ then
        BREAK
    end if
end for
```

$\nabla^*_{\mathcal{G}}(\cdot)$, which is the adjoint operator (divergence) of $\nabla_{\mathcal{G}}$ and τ is the spectral norm (maximum eigenvalue) of $\nabla_{\mathcal{G}}$, i.e, $\tau = \|\nabla_{\mathcal{G}}\|_2$. The algorithm is a simple application of the proximal splitting methods commonly used in signal processing [57]. Note that the solution to the ℓ_1 norm is a simple element-wise soft-thresholding operation. Although, the optimization problem here has been solved using the forward-backward primal dual [57] available from Open Source tool-box UnLocBox [58] it may be solved using a variety of other optimization tools.

2.5 Reconstruction Phase

In order to show the effectiveness of the proposed approach, the refined sinogram was reconstructed using different analytical and iterative methods. Analytical methods such as FBP [59] are the most commonly used methods employed for ET, CT scanners and other tomographic modalities since they are relatively simple and computationally trivial but do not provide robustness to noise. Iterative methods can partially remedy this but with a varying degree of certainty and almost never perfectly. Iterative row action methods such as algebraic reconstruction technique (ART) [60] and simultaneous reconstruction technique (SIRT) [61] are commonly used due to their semi-convergence[2] and regularization capabilities. ART or Kaczmarz method operates on each row/equation of the linear system defined in section I and treats it as a hyperplane in vector space. Starting from an initial guess each iteration entails sweeping all rows by taking orthogonal projections successively until it converges to a solution. This can be defined as,

$$x^{k+1} = x^k + \lambda \frac{b_i - a_i^T x^k}{\|a_i\|_2^2} a_i \qquad \lambda \in (0, 2), \tag{2.4}$$

where k is the iteration, λ is the relaxation parameter[3] and the rest of the vari-

[2] In some literature the concept of semi-convergence is also referred to as ‘early stopping’. This has been clearly explained in [18, 31]. Some litrature also attributes ’early stopping’ as a semi-convergence-based stopping criteria for algebraic methods.

[3] The relaxation parameter should not be confused with the regularization parameter in variational

ables are the same as mentioned in the preliminaries. The fact that during the early iterations, Kaczmarz quickly semi-converges to an approximate solution, makes it an ideal method to test sinogram denoising. A refined sinogram should semi-converge more quickly to a smaller error and should diverge more slowly from the most accurate solution. SIRT methods access all rows simultaneously and generally have better regularization capabilities, compared to ART methods. Although there are several different variants of SIRT, in this study Cimmino's method [62] was used which can be defined as follows:

$$x^{k+1} = x^k + \lambda_k \frac{1}{m} \sum_{n=1}^{m} w_i \frac{b_i - \langle a_i, x^k \rangle}{\|a_i\|_2^2} a_i. \tag{2.5}$$

Cimmino's method usually converges faster than ART and has similar semi-convergance properties. However, these properties vary substantially based on the data being reconstructed and the amount of noise in the data. A more detailed account regarding algebric reconstruction method including ART, SIRT and their variants will be discussed in Chapter 4. Here they can be treated as standard reconstruction methods employed to test the effectiveness of sinogram denoising explained in the previous sections.

2.6 Experiments and Results

2.6.1 Simulation Setup

The Validation Problem

Validation is a serious problem not only in Cryo-ET or other tomographic modalities but with all inverse problems in general. Most iterative image reconstruction (IIR) methods lack proper validation. Such literature is particularly deficient for Cryo-ET where FBP, SIRT and their variants have been the standard methods for a long time and remain the most commonly used methods today. In order for a new method to justify superiority over existing methods simply showing the ability to reconstruct cleaner images is not sufficient. Such a method must also justify that it does not produce superfluous artifacts. Another problem is that the notion of resolution is often unclear and misunderstood, there are no theoretical results that give bounds to the best possible[4] resolution [12]. Validation in Cryo-ET and most other heavily *ill-*

regularization problems. The relaxation parameter is merely the step size algebraic methods take during each iteration to approach an approximate solution. The regularization parameter on the other hand is what determines a balance between the regularity of the solution and its faithfulness to raw data. In the case of algebraic methods, their regularization properties are purely associated with semi-convergence to an approximate solution. Some literature such as [18] suggests that the iteration number for algebraic methods can be associated with the regularization parameter however this is debated. Although, the semi-convergance properties of SIRT and its variants have been studied extensively such analysis for ART remains an open problem.

[4] Not to be confused with 'ideal' or 'theoretical' resolution which is given by the Crowther's Criteria explained in Equation 1.1.

posed[5] *inverse problems* must be using simulated phantoms, *i.e.*, images, datasets and specimens whose ‘true’ values (intensities, frequency analysis etc.) are already known. In order to minimize the possibility of committing the so-called "inverse crime" [63, 64], a new method should be able to recover several different simulated phantoms and the results should be compared with the ground truth[6] and the recovered resolution should be analyzed using frequency analysis tools. Simulating a phantom to generate a tilt series close to actual data acquisition settings is not trivial and has been discussed in detail in [65].

Analyzing phantoms reconstructed from various methods is also a non-trivial problem and the effectiveness of various established methods is debated. Standard signal-to-noise ratio and ℓ_2 residual norm error-based metric can compare the original phantom with the reconstructed data but fails to effectively analyze the amount of information compromised during reconstruction. Due to this reason besides using standard ℓ_2 residual norm errors, intensity analysis and visual inspection of reconstructed phantoms frequency analysis based on radially averaged power spectrum was also used. In certain cases Fourier ring correlation[7] was also used. The choice of these metrics depends on the nature and objective of the experiment and has been explained as they are used in the proceeding subsections and subsequent chapters. As the objective of this thesis is more relevant to image reconstruction methodology rather than determining structures of molecules we will stick to phantom based studies with some tests on real data.

Experiment 2.1: Denoising and Reconstructing a Shepp-Logan Sinogram using ART (Kaczmarz)

Aim

To investigate the extent of improvement that can be achieved by denoising a sinogram using graph-based Total Variation and reconstructing the phantom using ART (Kaczmarz Method) and FBP.

Method

Raw data was simulated from a 64 × 64 Shepp-Logan [26] phantom by collecting 36 equally spaced projections from the data (*i.e.*, 0:5:180). Sinogram, S, is of size 95×36[8] where each column corresponds to projections at each of the equally spaced 36 angles from 0 to 180 degrees. Random Gaussian noise with mean 0 and variance adjusted

[5] The term "heavily" ill-posed inverse problems was coined to distinguish Cryo-ET from other problems which are ill-posed, *i.e.*, lack uniqueness, stability and sufficient data but can be estimated or solved with greater precision. This terminology is simply a testament to the fact that ’not all ill-posed inverse problems are created equal’ and is inspired by [12] and [18].

[6] The term *"Ground Truth"* has a variety of meanings in various fields and can be a confusing for interdisciplinary readers; in this thesis it simply means the ’real true’ value. Recent machine learning literature relates the word with ’data that is known to be correct’.

[7] Fourier Ring Correlation is a 2D form of Fourier Shell Correlation commonly used in structural biology to determine the resolution of structures.

[8] It should be noted that the size of the sinogram here is corrosponding to the *radon*(x, θ) function in MATLAB 2014a and assumes a rectangular matrix input.

Table 2.1: Comparison of Regular and Graph Denoised (GD) Reconstructions

Phantom	FBP	FBP-GD*	ART[1]	ART-GD*
Shepp-Logan (RN=0.05)	6.41	6.36	4.58	3.89
Shepp-Logan (RN=0.08)	6.76	6.53	5.53	4.44
Phantom	FBP	FBP-GD*	SIRT[2]	SIRT-GD*
Smooth (RN=0.05)	8.51	3.68	4.82	2.81
Smooth (RN=0.08)	13.16	4.83	6.65	3.82

* Proposed Method

to 8% of the norm is added to individual projections. For the graph based total variation denoising stage of the experiments, the sinogram is divided into $95 \times 36 = 3420$ overlapping patches of size 3×3. Larger patches tend to decrease the likelihood of finding similar regions and are generally not used. The graph $\mathcal{G}$ is constructed between the 3420 patches with the 10 nearest neighbors ($K = 10$) and σ for the weight matrix (Section 2.4.1) is set to the average distance of the 10-nearest neighbors. Various values of γ in the range of $[0, 10]$ are tested for denoising (Algorithm 1). The parameters are tuned by defining a logarithmic range followed by a linear range within the region of interest determined by the logarithmic range. The best γ is selected based on the minimum ℓ_2 reconstruction for the phantom. The relaxation parameter λ was also tuned to achieve best semi-convergence results. The relaxation parameter for ART and SIRT is tunes tuned between zero and two and was tuned via tools available in the AIR tools package [31].

Conclusions

Fig. 2.4 presents the reconstructed phantom using FBP and ART with both the regular sinogram as well as the graph denoised (GD) sinogram. ℓ_2 reconstruction error variations with the iterations, k, have been shown in Fig.2.5. ℓ_2 reconstruction error for the k^{th} iteration relative to the original phantom was calculated using $\left\|x^k - x^*\right\|_2$, where x^* is the original phantom and x^k is the result of the k^{th} iteration of the iterative method being used. A close analysis of the results shows that graph-based denoising helps to attain lower reconstruction error for analytical (FBP) as well as iterative (ART and SIRT) methods. This result is also visually obvious from the reconstructed phantoms. Error curves and intensity profiles show that for both phantoms, a smaller error can be achieved using graph denoised sinograms rather than regular raw data. Shepp-Logan reconstruction with ART specifically shows that semi-convergence to an approximate solution is faster and divergence from this approximation is slower, an indication that the system has lower noise. These results show that the proposed method is extremely general and can be adapted for any tomographic reconstruction modality, regardless of the reconstruction method employed.

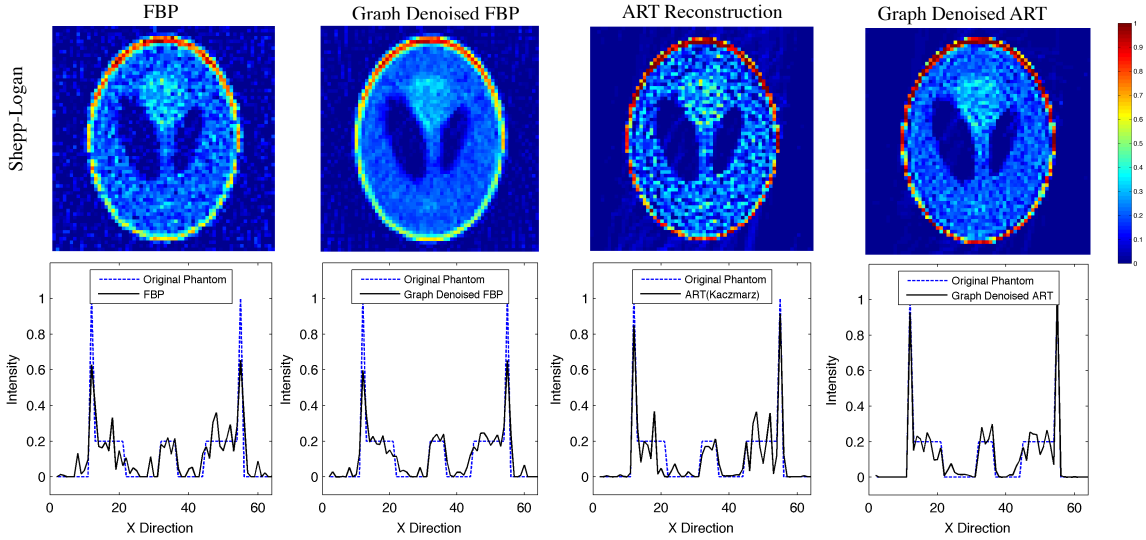


Figure 2.4: Recovered Shepp-Logan phantoms and intensity profiles for the following methods: 1) Filtered back projection (FBP), linearly interpolated, Ram-Lak filtered. 2) Graph denoised FBP (FBP-GD). 3) ART reconstruction (without denoising) and 4) Graph denoised ART / SIRT reconstruction (ART-GD, SIRT-GD). Images can be best viewed in color in the electronic version of the chapter.

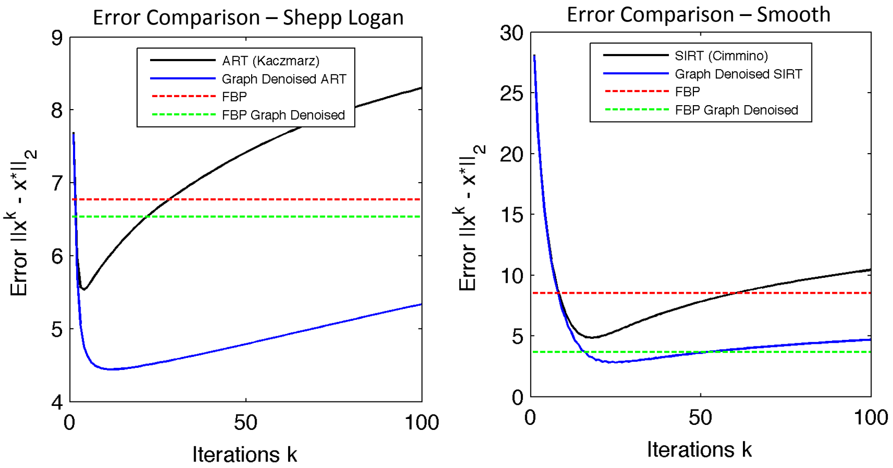


Figure 2.5: ℓ_2 reconstruction error variations with k iterations, for the following methods: 1) Filtered back projection (FBP), linearly interpolated, ram-lak filtered. 2) Graph denoised FBP (FBP-GD). 3) ART reconstruction (without denoising) and 4) Graph denoised ART reconstruction (ART-GD, SIRT-GD).

Experiment 2.2: Denoising and Reconstructing a Smooth Sinogram using SIRT (Cimmino)

Aim

To demonstrate the effectiveness of sinogram denoising using a smooth phantom reconstructed using Cimmio's Method.

Method

Raw data was simulated from a 64×64 Smooth [31] phantom by collecting 36 equally spaced projections from the data (*i.e.*, 0:5:180). Sinograms, S, is of size 95×36 where each column corresponds to projections at each of the equally spaced 36 angles from 0 to 180 degrees. Random Gaussian noise with mean 0 and variance adjusted to 5% of the norm of the individual projections is added to individual columns of S. The remaining simulation setup is similar to experiment 2.1.

Conclusions

Fig. 2.6 shows the reconstructed smooth phantom as well as the intensity profiles. The semi-convergance behaviour of SIRT can be observed from Fig. 2.5. Clearly there is an improvement in the signal, it will be further investigated in the next experiment that how much frequency information is compromised to achieve this improvement. The results presented in this experiment and experiment 2.1 are from relatively less amount of data (10% maximum possible data, given one projection can be collected every degree). However, these do not fully simulate the cryo-ET situation and without a detailed frequency analysis it is not possible to deduce the effectiveness of the method.

Experiment 2.3: Effect of Graph Sinogram Denoising on the Missing Wedge using a Ring Phantom (More Realistic Cryo-ET Simulation Setting)

Aim

To investigate the extent of improvement that can be achieved by denoising a sinogram with a missing wedge using graph-based Total Variation and reconstructing the phantom using various reconstruction methods. Also, analyzing the resolution compromised as the result of using graph-denoised sinograms.

Method

2D Cryo-ET Phantom Simulation Model

Experiments 2.1 and 2.2 and the results described therein simply give a feel for the method. In oder to avoid unrealistically optimistic results from the proposed method further tests were conducted with more realistic simulations taking into account further challenges associated with Cryo-ET. In this experiment we generate the angles according to the missing wedge (*i.e.*,$-70° \leq \theta \leq +70°$). Moreover, even though there has been significant development in alignment procedures for Cryo-ET tilt series, the data is always misaligned to a certain degree. To emulate this effect and ensure a more

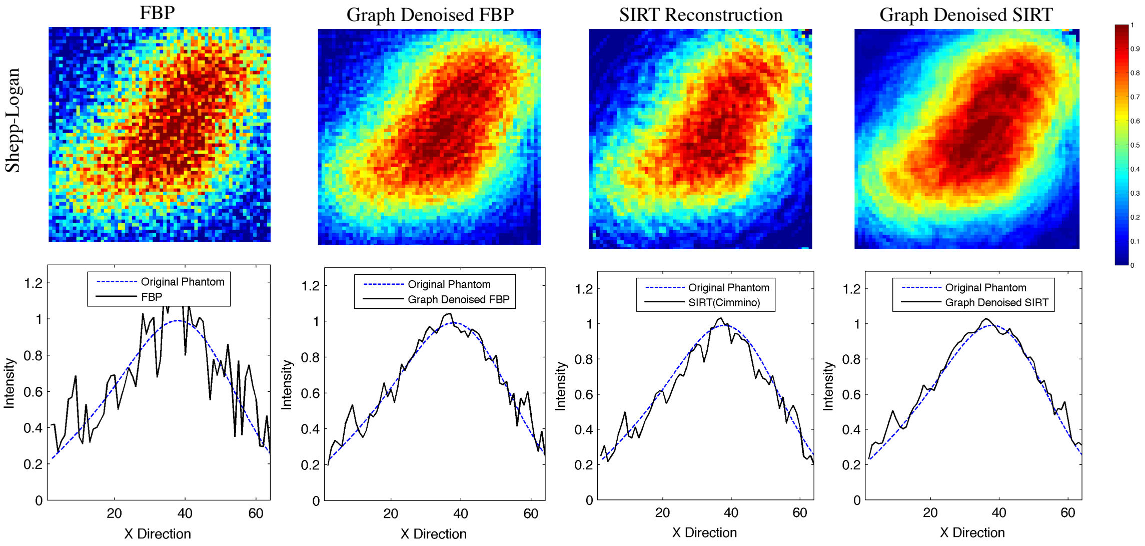


Figure 2.6: ℓ_2 reconstruction error variations with k iterations, recovered phantoms and their intensity profiles for the following methods: 1) Filtered back projection (FBP) 2) Graph denoised FBP (FBP-GD) 3) ART / SIRT reconstruction (without denoising) and 4) Graph denoised ART / SIRT reconstruction (ART-GD, SIRT-GD). Images can be best viewed in color in the electronic version of the chapter.

realistic simulation setup the projections were automatically shifted to a randomly scaled amount within a ±1 pixel margin. This simulation setup was inspired by [66] and was implemented using MATLAB. In order to avoid the 'inverse crime' of using the same model or discretization system to create and reconstruct the phantom any discretization steps were completed in Mathematica. This was only rarely the case since most phantoms used were already discretized and generated by a system independent of MATLAB. This simulation setup is used throughout this chapter as well as chapter 3.

Raw data was simulated from a 128×128 Ring phantom [66] by collecting 71 equally spaced projections from the data (*i.e.*, $-70° \leq \theta \leq +70°$, every two degrees). Sinograms, S, is of size 181×71 where each column corresponds to projections at each of the equally spaced 71 angles with a missing wedge in between (the effect of the missing wedge was demonstrated in Chapter 1). A mixture of Random Gaussian and Poisson noise was added to the projections (hereafter this combination will simply be referred to as noise). For the graph based total variation denoising stage of the experiments, the sinogram is divided into $181 \times 71 = 12851$ overlapping patches of size 3×3. The graph $\mathcal{G}$ is constructed between the 12851 patches with the 10 nearest neighbors ($K = 10$) and σ for the weight matrix (Section 2.4.1) is set to the average distance of the 10-nearest neighbors. Various values of γ in the range of $[0, 10]$ are tested for denoising (Algorithm 1). The best γ is selected based on the minimum ℓ_2 reconstruction for the phantom. The denoised sinogram was reconstructed using Kaczmarz, Cimmino's Method, the relaxation parameter λ was also tuned to achieve best semi-convergence results. FBP reconstructions were linearly interpolated and ram-lak filtered.

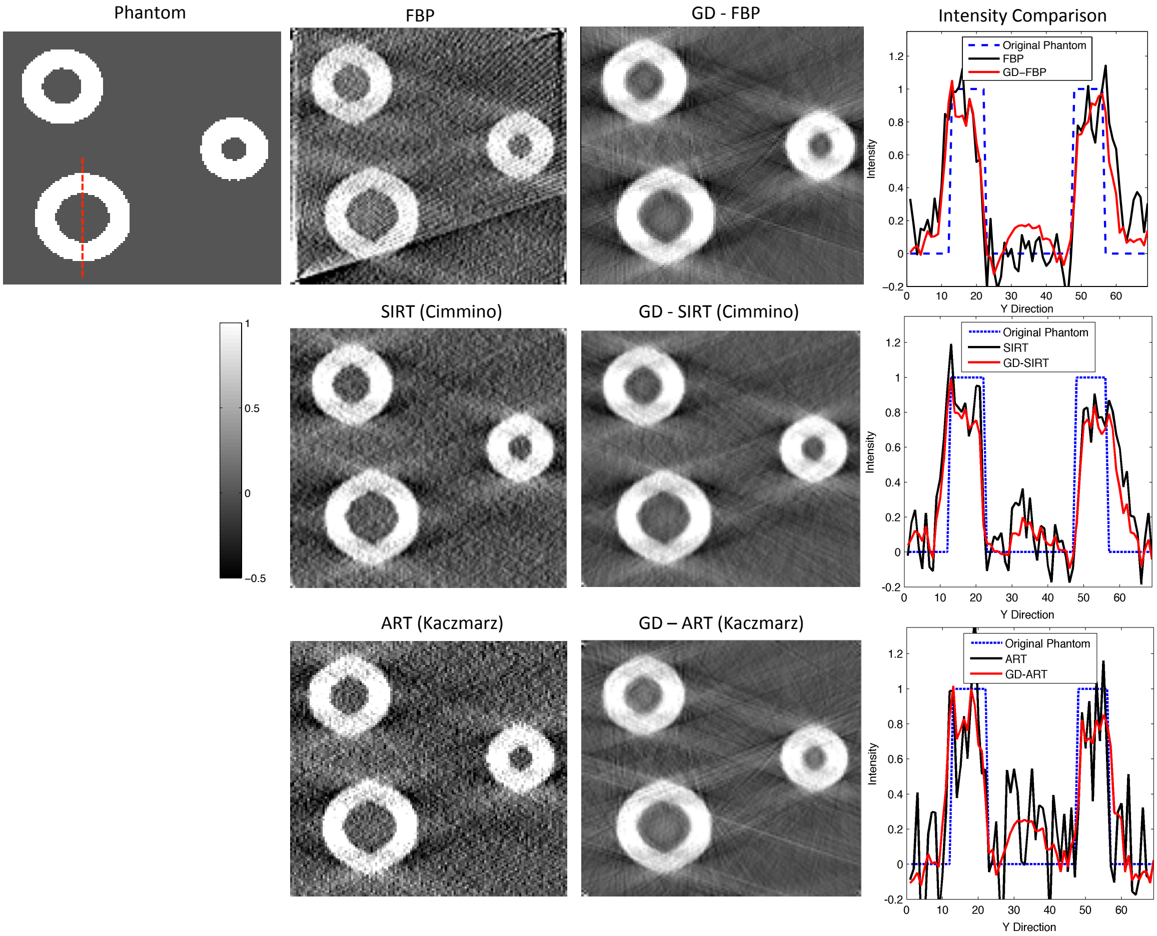


Figure 2.7: The phantom, reconstructions and intensity profiles of reconstructing a 128 × 128 ring phantom from 71 equally spaced projections with a missing wedge ($-70^\circ \leq \theta \leq +70^\circ$) in the middle. It can be clearly seen that the effect of missing wedge elongation is prominent in the graph-denoised reconstructions. However, it can be seem from the intensity profiles that the overall noise in the reconstruction is much more constrained.

Conclusions

Fig. 2.7 shows the reconstructions and intensity profiles of the reconstructed ring phantom. The reconstructions from the graph denoised sinogram are much more cleaner and the intensity profiles are more constrained. In order to investigate the amount of high frequency details compromised a radially averaged power spectrum[9] was calculated and the results are shown in Fig. 2.8. It can be clearly seen from both the reconstructed phantoms as well as from the RAPS that due to the missing wedge elongation and missing data effects the edges of the circles are elongated. This shows up as high frequency details in some methods. The graph denoised reconstructions are much more constrained. The RAPS in general shows that all reconstruction methods employed are able to deduce the basic structure of the phantom being reconstructed given the fact that for lower wave numbers the spectrum is very similar. For higher frequency FBP in specific adds superfluous details in the reconstruction as seen from the RAPS its spectrum reaches beyond the ground truth. All reconstructions from the graph denoised sinogram are closer to the ground truth demonstrating that although there is some loss in frequency the lost component may be from superfluous higher frequency details and noise. That said, the proposed method does not guarantee perfect recovery and the final results will always depend on the type of data being reconstructed and the amount of noise present in the data. It should be noted that given the fact that the sinogram is being denoised all reconstructions give a better ℓ_2 error rate except for situations when γ is too high that it smooths out all useful information from the sinogram. For this reason ℓ_2 errors have not been used as a metric for this and proceeding experiments in this chapter. However, they were used as a stopping criteria for some algebraic reconstructions as well as for tuning the regularization and relaxation parameters[10]

Experiment 2.4: Effect of Graph Sinogram Denoising on the Missing Edge using a 256×256 Shepp-Logan Phantom

Aim

To investigate the extent of improvement that can be achieved by denoising a sinogram corresponding to a 256×256 Shepp-Logan Phantom with a missing wedge using graph-based Total Variation using ART (Kaczmarz Method) and FBP.

Method

The methodological setup is exactly the same as experiment 2.3 except there is 7% noise added to the sinogram. All parameters were tuned to achieve best results.

[9] The radially averaged power spectrum (RAPS) is the mean spectrum independent of direction in the image. More specifically, it is the average of all possible directional power spectra.The metric provides an easy way to compare and contrast information contained in 2D spectra in 1D.

[10] An interesting future direction is to choose the regularization parameter (γ) of the sinogram denoising and the relaxation parameter for algebraic methods (λ) based on both the ℓ_2 error norm as well as RAPS. Choice of these parameters is an open research problem and it is non-trivial to find rules that can be generically applied to all types of data.

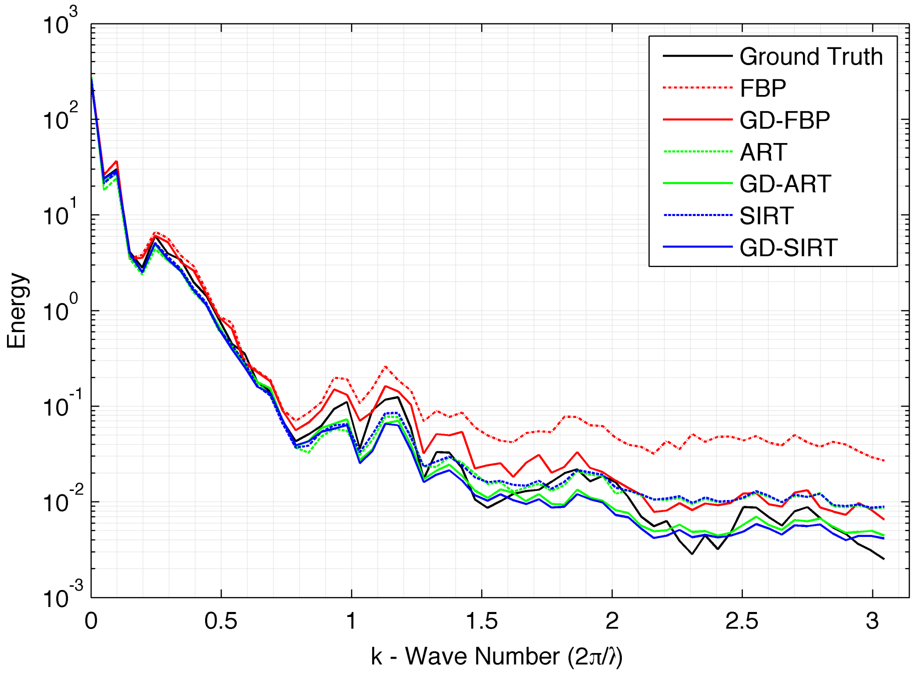


Figure 2.8: Radially averaged power spectrum (RAPS) of ring phantom reconstructions using FBP(linearly interpolated, ram-lak filtered), FBP using GD-Sinogram ART (Kaczmarz, SC=100), ART using GD-Sinogram, SIRT(Cimmino, SC=100) and SIRT using GD-Sinogram. This clearly shows that all methods are able to determine the basic structure of the phantom as seen from the lower wave number values. Due to the missing wedge and noise some methods add extra high resolution details to the result which are erroneous and incorrect. These erroneous high resolution details can be reduced by non-locally denoising the sinogram using the proposed method.

Conclusion

Fig. 2.9 shows the reconstructed phantoms and intensity profiles. The missing wedge elongation is certainly much larger in FBP but this can be reduced to some extent with standard algebraic methods even when GD sinogram is not employed. The graph denoised reconstructions are much cleaner and have constrained amount of noise. Fig. 2.10 shows a particularly interesting behavior for FBP, regular reconstruction in this case is adding too much erroneous high frequency details in the reconstructed image. It should be noted that this is not always the case, and depending on the randomized noise added the reconstruction may not add too many high frequency details which are not true. However, this effect becomes more prominent with missing data due to streaking artifacts and the missing wedge elongation. On using FBP with the graph-denoised sinogram it becomes over-smoothed and loses most high frequency details. This is remedied by using iterative ART or SIRT. It should be noted that both GD-ART and GD-SIRT stay below the ground truth, *i.e.*, they possibly don't include high frequency details which are erroneous but do lose resolution while covering up for those details. That said, some amount of resolution loss is expected from all denoising methods. It has been shown in previous literature that non-local graph-based methods preserve higher frequency details better than local denoising methods, thus a direct comparison between local and non-local sinogram denoising was tested but not included in this chapter.

Experiment 2.5: Results with Real Data

Aim

To investigate the extent of improvement that can be achieved by denoising a sinogram when reconstructing real Cryo-ET data from a biological specimen.[11]

Method

A 256×256 cryo-ET image from open-sourse EMD database was emulated to generate 141 equally spaced projections with a missing wedge in between, the 1D aligned projections had noise which is naturally added during data collection hurdles mentioned in Chapter 1. A sinogram was constructed and denoised using graph total variation according to the setup explained earlier while tuning the regularization parameter, γ. The projections were reconstructed using FBP and GD-FBP and intensity profiles and RAPS was analyzed.

[11] This thesis focuses on image reconstruction from a methodological prospective and testing of these methods on real data and to answer specific biological questions is beyond the scope of this thesis. This is basically because these methods are evolving and the focus of these methods is to improve image reconstruction methodology which is much more evident from Chapter 3. These results are presented as a teaser for how this method can behave in a real data setting, to achieve a better result more investigation in necessary in terms of parameter tuning and reducing the complexity of graph construction. The computational complexity associated with graph construction is $\mathcal{O}(n^3)$ which can be significant for large datasets.

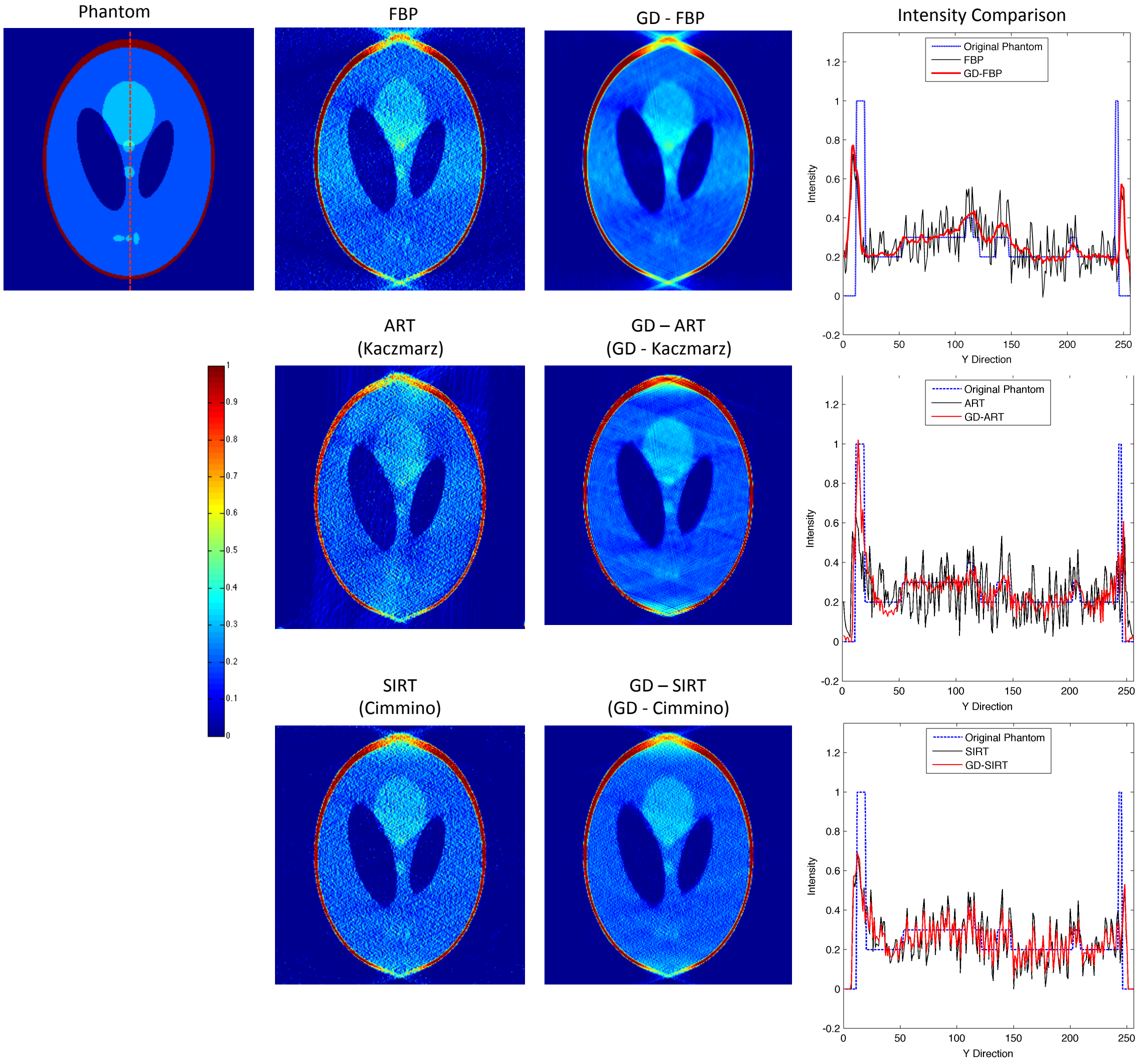


Figure 2.9: The phantom, reconstructions and intensity profiles of reconstructing a 256×256 Shepp-Logan phantom from 71 equally spaced projections with a missing wedge ($-70° \leq \theta \leq +70°$) in the middle. It can be clearly seen that the effect of missing wedge elongation is prominent in the graph-denoised reconstructions. However, it can be seen from the intensity profiles that the overall noise in the reconstruction is much more constrained.

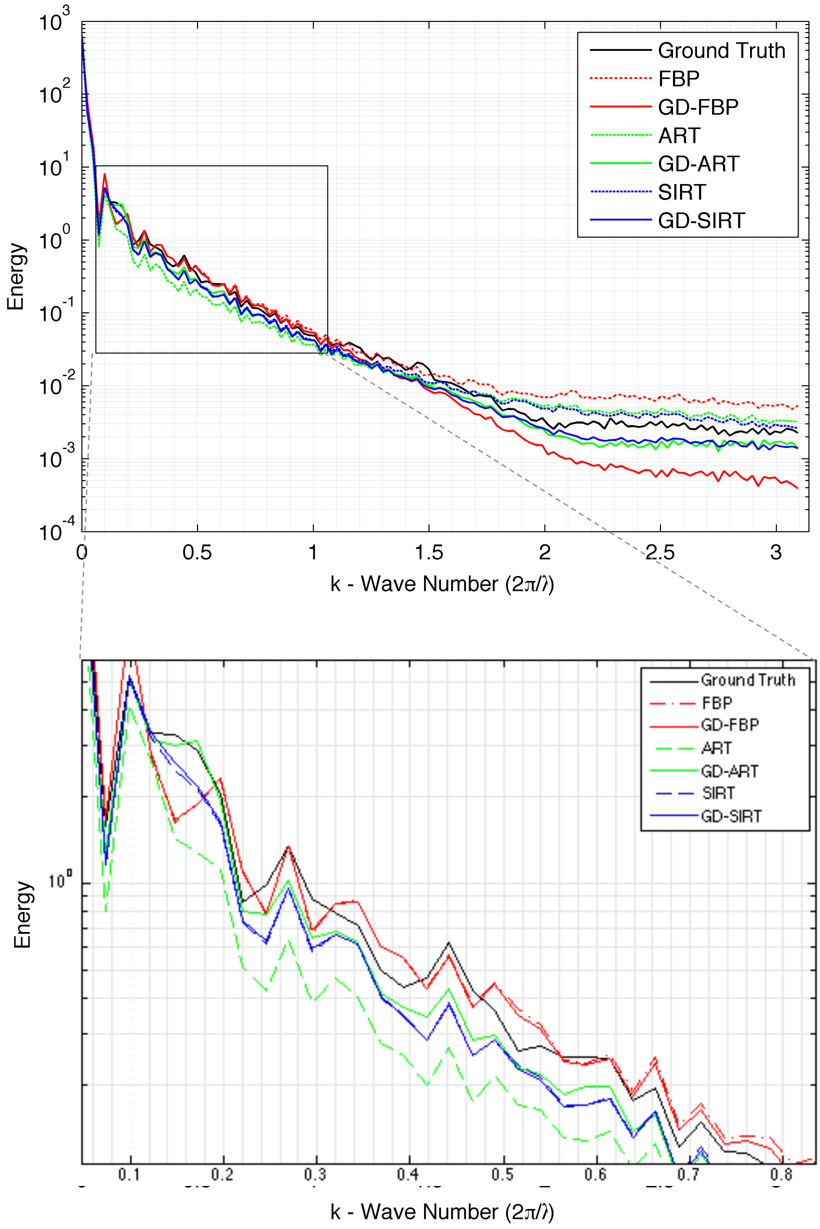


Figure 2.10: RAPS of the Shepp-Logan phantom reconstructions using FBP(linearly interpolated, ram-lak filtered), FBP using GD-Sinogram ART (Kaczmarz, SC=100), ART using GD-Sinogram, SIRT(Cimmino, SC=100) and SIRT using GD-Sinogram. This clearly shows that all methods are able to determine the basic structure of the phantom as seen from the lower wave number values. Due to the missing wedge and noise some methods add extra high resolution details to the result which are erroneous and incorrect. These erroneous high resolution details can be reduced by non-locally denoising the sinogram using the proposed method.

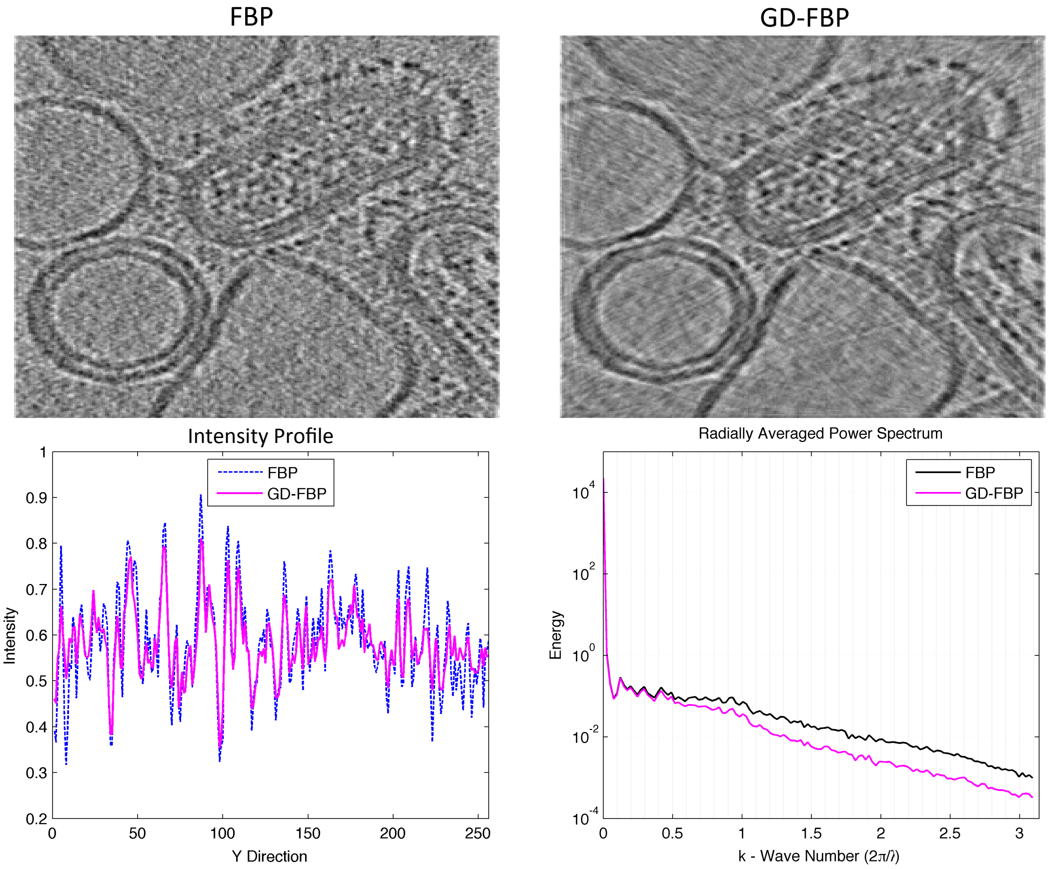


Figure 2.11: Figure showing the FBP and GD-FBP reconstruction of Cryo-ET emulated data as well as their RAPS and intensity profiles. The data used is from EMD-4067. It can be seen that although the image reconstructed from the graph-denoised image is smoother it loses significant frequency details. There high frequency details could also be from noise. The image was reconstructed from 141 1D projections. Data used is from an Influenza virus available from opensource EM-Data-Base Entry: 4067.

Conclusions

Fig. 2.11 shows reconstructions, intensity profiles and the RAPS. It can clearly be seen that the FBP reconstruction has a higher RAPS as compared to the GD-FBP reconstruction. This makes sense since the sinogram is denoised. Although FBP has higher frequency details as seen from simulated data in the previous experiments it is evident that these details can come from erroneous details due to missing data and missing wedge elongation which are picked up by the method. The intensity profile shows recovery of the signal while loosing some high intensity information.

2.7 Limitations and Shortcomings

The major limitation of this work is the fact that although non-local TV can perform better than local TV and filtering based methods to denoise the raw data it does not come with a guarantee of signal recovery without loosing high frequency details in the image. Another constraint is that the current method needs to build a nearest neighbor similarity patch graph on overlapping patches. This process has a complexity which scales with $\mathcal{O}(n^3)$. Although graph construction can be approximated using procedures discussed in the proceeding chapter, the approximation procedure should not be preferred for sinogram denoising. Graph approximation is fine when graphs are being used to promote sparsity in compressed sensing type reconstructions since the error originating from the approximation will have minimal effect on the final results. While manipulating raw data this type of approximation may generate minor denoising errors that can become arbitrarily large during the reconstruction process.

Another factor contributing to the complexity of the method is that for real 3D data the degree of similarity has to be associated with all 2D projections, this can be achieved by tensor processing where a cube instead of a patch graph is built and a weight matrix is constructed on overlapping cubes. This further contributes to the complexity. Given the fact that tensor-based non-local processing is still an active research area and procedures of complex multidimensional non-local processing are still under development further contributes to the problem. An approach to remedy this is to rearrange the data in a way that each 2D matrix of the 3D sinogram has a column or row of pixels originating from the same column or row in the 2D projections. This can be proceeded by denoising each 2D layer of the 3D sinogram non-locally slice by slice. Although this method could produce interesting results it will fail to associate various regions of the sinogram and the risk of artifacts would be high.

Another major shortcoming is the tuning for the regularization parameter. Selection of such parameters in an open research problem not only for Cryo-ET but for inverse problems in general. In this case the selection is more critical because we are manipulating the raw data, in the case of variational regularization based reconstruction methods where the fit and the regularity of the solution is being balanced iteratively, this is much less of a problem. Choosing a high regularization parameter can overly smooth the sinogram whereas choosing a small parameter can increase the noise in the reconstruction. The solution to this is testing each problem with all possible parameters and estimating the parameter for similar problems. Moreover, as seen from the results, the proposed method does not have the ability to recover missing data. The effects seen on the missing wedge are merely the effect of smoothing.

2.8 Conclusions

This chapter presented a sinogram denoising preprocessing method based on non-local graph denoising. The method is interesting in the sense that it makes use of the fact that unlike the reconstructed image the sinogram has a piecewise smooth structure because of the high degree of redundancy in the data. The effectiveness of the method was shown with a variety of phantoms in several different conditions as well as some real data. Although, the method can denoise the reconstructions it does not have the ability to reconstruct missing data. This study is proceeded by a more complex non-local simultaneous denoising and reconstruction method in the next chapter which makes use of modern sparsity exploiting image reconstruction and shows how they can be used to partially recover missing data.

Chapter 3

Adaptive Graph Total Variation for Tomographic Reconstructions

3.1 Chapter Outline

Sparsity exploiting image reconstruction (SEIR) methods such as compressed sensing (CS) have been extensively used with Total Variation (TV) regularization for a variety of inverse problems including tomographic reconstructions. Over the past few years the interest in CS has exponentially grown due to the ability of such methods to reconstruct from limited data. Local TV regularization methods fail to preserve texture details and often create additional artifacts due to over-smoothing. Non-Local TV (NLTV) has been recently proposed as a solution to this but lacks continuous update and becomes computationally complex if updated every iteration. In this chapter we propose Adaptive Graph-based TV (AGTV). Similar to NLTV our proposed method goes beyond spatial similarity between different regions of an image being reconstructed by establishing a connection between similar regions in the image regardless of spatial distance. However, it is computationally more efficient and involves updating the graph prior during every iteration making the connection between similar regions stronger. Moreover, it promotes sparsity in the wavelet and graph gradient domains. Since TV is a special case of graph TV the proposed method can also be seen as a generalization of SEIR and TV methods[1]. Extensive experimentation shows that when compared to other state-of-the-art methods the proposed method achieves a better result in terms of reconstructing from missing, noisy and erroneous data while preserving high frequency components of the signal. This work has been archived as [67].

3.2 The Story So Far

Chapter one introduced the fundamental concepts, challenges and problems associated with Cryo-ET in specific and tomography in general. This was proceeded by Chapter 2 which introduces and analyses a pre-processing-based method to denoise the raw data

[1] *Dr.Bo Zhao* of Harvard University deserves to be acknowledged for pointing this interesting aspect of the method at the 38th IEEE Engineering in Medicine & Biology Conference, Orlando, FL. Aug, 2016.

before it is reconstructed. Although, graph-based non-local denoising works better than standard SIRT and ART reconstructions it still manipulates the raw data which comes with a risk of loosing useful information while smoothing the raw sinogram.

As mentioned in Chapter one tomographic modalities such as electron tomography (ET) and computed tomography (CT) suffer from low-dose and missing data constraints which lead to noisy and erroneous reconstructions. In the case of ET, biological samples are rotated in a transmission electron microscope (TEM) and cannot be exposed to high levels of electron dose because it leads to the degradation of the sample [4, 9]. Whereas, in the case of CT low-dose has been a clinical objective to prevent the patient from over exposure to ionizing radiation [68–70]. One way to reduce the dose is by reducing the number of projections or views collected during tomographic imaging. However, this leads to missing data, which results in reconstructions effected by noise. In short, like several other tomographic modalities the Cryo-ET reconstruction problem is made tedious by the fact that the problem is ill-posed due to under sampled, missing and noisy raw data and has no unique solution. In other words the system of linear equations being solved is highly under-determined, making signal recovery difficult.

A variety of analytical [21, 37, 71] and iterative image reconstruction [7, 15, 39, 72–74] algorithms have been employed over the years to deal with the low-dose and missing data problem. Iterative image reconstruction (IIR) methods traditionally yield more favourable results when reconstructing noisy and missing data due to their regularization properties and ability to incorporate prior knowledge. Algebric IIRs such as ART, SIRT and their variants have mild regularization capabilities steming from their semi-convergance properties [18, 31]. Recent statistical approaches to IIRs are often based on the *Ordered Subset* (OS) technique [39, 75–77].

3.3 Sparsity Exploiting Image Reconstruction (SEIR)

Over the recent years the concepts of recovering a signal from under-sampled raw data have undergone a major paradigm shift. This is due to the development of a new class of algorithims known as sparsity exploiting methods (SEMs) with compressed sensing[2] [78, 79] at its core. It has been well known that prior knowledge can improve signal recovery for tomographic reconstructions. This prior knowledge can be as little as knowing that the outcome is non-negative. Compressed Sensing exploits the fact that most signals are sparse or compressible in some known transform domain. However, for realistic applications CS alone is not sufficient to cater for the sparsity of the Gradient Magnitude Image (GMI), therefore, it is used along with Total Variation (TV) in CT and ET [79–84]. The joint CS and TV setup is referred to as CSTV in the sequel. The authors of [66] use CSTV for ET and show that the reconstruction error is better than other iterative algorithms when reconstructing from limited data. A comparison of statistical iterative reconstruction methods with CSTV has been provided in [83]. An analysis of the CSTV method for reconstructing needle shaped biological specimens is

[2] Also known as compressive sensing, compressive sampling. The label, “Compressed Sensing(CS)” has been used in a variety of different ways in various literature. Although, initial literature suggests that CS should be associated with methods where these is an absolute guarantee of signal recovery, there is a plethora of literature where CS simply means sparsity exploiting reconstruction.

presented in [17]. CSTV has also been used in photo-acoustic tomography as presented by the authors of [85]. A concise review of the iterative reconstruction algorithms, including TV regularization methods is presented in [43].

A more advanced type of TV regularizer, known as the non-local TV (NLTV) [86] has been shown to be much more efficient for inverse problems, such as denoising, inpainting and super-resolution [52, 87]. In contrast to simple TV, which takes into account the similarity of a region with only its neighboring regions, NLTV overcomes this limitation by associating a similarity measure of every region of an image with all other regions (full NLTV) or a few regions in the spatial neighborhood (partial NLTV). This will be explained in more detail in the upcoming sections of this chapter. As an application of NLTV, the authors of [88] presented a reconstruction method for Magnetic Resonance Imaging (MRI) via CS. NLTV has been further explored in a spectral CT setting in [89, 90].

A primary short-coming of full NLTV-based methods used in CT, ET and MRI settings is that the similarity matrix constructed in the beginning from the initial estimate of the sample is not updated throughout the algorithm, for example [88]. The authors of [89] construct a similarity matrix from the solution of a TV based minimization method and then keep it fixed throughout the algorithm. A primary reason for keeping it fixed is that the NLTV based method suffers from a high cost of associating a similarity measure between every pair of regions in an image / sample. For an image of the size $n \times n$, NLTV costs n^4 and is computationally cumbersome for high resolution applications. Furthermore, it does not make sense for every region to be connected to all other regions. Therefore, NLTV requires a threshold parameter which, based on the euclidean distance can decide if the similarity is strong enough to be non-zero. However, there is no appropriate method to fix this parameter and it depends on the scale of pairwise distances. Although, the results obtained by the above methods are state-of-the-art, the final reconstruction would be more faithful to the original one if one updates the similarity matrix from the simultaneously reconstructed sample regularly throughout the algorithm.

Contributions

In this chapter Compressed Sensing and Adaptive Graph Total Variation (AGTV) is proposed as a method for simultaneous reconstruction and denoising of tomographic data. The method uses a more sophisticated and scalable form of NLTV and has a much lower computational complexity as compared to NLTV[3]. The proposed method promotes the use of $\mathcal{K}$-nearest neighbor graphs in Graph Total Variation (GTV), where $\mathcal{K}$ is fixed and unlike NLTV, does not depend on the scale of the pairwise distances. The graph is constructed by using an approximate nearest neighbor search algorithm (FLANN) [91]. Furthermore, our method requires updating the GTV prior in every

[3] The proposed method is computationally less complex when compared to an adaptive version of full NLTV. Generally, NLTV in literature is non-adaptive and keeps the similarity matrix constant. The difference between AGTV and NLTV has to be clarified, the proposed method AGTV is a CS-type sparsity exploiting reconstruction method which uses graph-based non local TV as an additional sparsifying transform. NLTV in the context of this chapter it means a similar setup to AGTV where NLTV is used as an additional sparcifying transform

iteration by constructing a new graph from simultaneously reconstructed sample.

3.4 Compressed Sensing & Adaptive Graph Total Variation

Our proposed method involves simultaneous denoising and reconstruction of tomographic projections and constitutes the following important components:

1. *Compressed sensing type sparse reconstruction.*

2. *Adaptive Graph total variation regularizer for improved denoised reconstruction.*

3.4.1 Optimization Problem

We first present the optimization problem under consideration and then study each of the terms of the objective function in detail. Let $S \in \Re^{p \times q}$ be the sinogram corresponding to the projections of the sample $X \in \Re^{n \times n}$ being imaged, where p is the number of rays passing through X and q is the number of angular variations at which X has been imaged. Let $b \in \Re^{pq}$ be the vectorized measurements or projections ($b = vec(S)$), where $vec(\cdot)$ denotes the vectorization operation and $A \in \Re^{pq \times n^2}$ be the sparse projection operator. Then, the goal in a typical tomographic reconstruction method is to recover the vectorized sample $x = vec(X)$ from the projections b. A highly underdetermined inverse problem needs to be solved for this type of reconstruction, which is even more tedious if the projections b are heavily corrupted with noise, measurement errors and missing data. To circumvent the problem of noisy projections, we propose

$$\min_{x} \|Ax - b\|_2^2 + \lambda \|\Phi^*(x)\|_1 + \gamma \|x\|_{gtv}, \tag{3.1}$$

where Φ is the wavelet operator and $\Phi^*(x)$, where $*$ represents the adjoint operation, denotes the wavelet transform of x and x_{gtv} denotes the total variation of x w.r.t graph $\mathcal{G}$. The first two terms of the objective function above comprise the compressed sensing based sparse reconstruction part of our method and the third term, to which we refer as the *graph total variation* (GTV) regularizer acts as an additional prior for denoising and smoothing the reconstructed sample. These two components are explained in detail below:

3.4.2 Compressed Sensing

Compressed Sensing (CS), introduced in the seminal papers of *Donoho* [78] and *Candes et. al.* [92] guarantees the recovery of a data sample from highly under-sampled measurements if the sample to be recovered can be sparsely represented in a basis. CS exploits fundamental principles of sparse approximation and transform coding that were established for image compression algorithms e.g. sparse representations from the discrete cosine transform (DCT) and the discrete wavelet transform (DWT) form the basis of the JPEG and JPEG-2000 image compression formats. Generally, for image compression after an image is sampled and its transform coefficients are calculated, smaller insignificant coefficients are discarded and fewer large components are stored. Since there are fewer large components, given the correct choice of the transform, the representation of the image can be considered as 'compressed'. Thus the major goal to compress an image is to represent it in as few coefficients as possible while still being able to reconstruct a reasonable quality image. In CS this concept of compressibility and transform sparsity is exploited during initial acquisition. The simplest way to understand CS is to think of the data acquisition process as a system of collecting a compressed form of data that has to be recovered. In this case the fundamental aim would be to record a small number of samples which best represent the important information in the signal such that they can be recovered perfectly. In the case of any tomographic experiment the measurements will be line integrals, and the sensing waveforms are the projected lines through the sample.

Let $x \in \mathbb{R}^{n^2}$ be the vectorized CT sample under consideration and $\Phi \in \mathbb{R}^{n^2 \times n^2}$ be the fourier basis, whose columns contain the fourier atoms. The proposed model, as described earlier, uses wavelets, however for the ease of description and explanation compressed sensing is explained with a Fourier matrix instead. For the case of Fourier matrix, $\Phi^*(x) = \Phi^\top x$ and one can sparsely represent the sample x in Φ as:

$$x = \sum_i \Phi_i c_i = \Phi c,$$

where $\|c\|_0 \ll n^2$, i.,e the number of non-zeros in c is much less than n. The sparse coefficients c are also referred to as the *transform codes*. Given this assumption, it is possible to estimate the unknown x from highly under-sampled measurements or projections $b \in \mathbb{R}^{pq}$, where $pq < n^2$. The projections or the measurements b can be obtained by applying a projection or sampling operator $A \in \mathbb{R}^{pq \times n^2}$ to the sample x. Thus, the projections b can be given as:

$$b = Ax = A\Phi c$$

Given b, it is possible to recover x by solving the following sparse recovery problem

$$\min_x \|\Phi^T x\|_1 \quad \text{s.t: } Ax = b,$$

where Φ^T denotes the forward transform (Fourier transform if Φ is a Fourier matrix). In case the projections or measurements b are contaminated with noise, one can solve the following recovery problem:

$$\min_x \|Ax - b\|_2^2 + \lambda \|\Phi^T x\|_1,$$

where λ is the regularization parameter which provides a trade-off between the sparsity and noise tolerance.

For tomographic applications, the projection matrix A is a line integral computed by rotating the sample x at different angles and the projections b are typically corrupted by Poisson noise. Therefore, CS is not only used as a reconstruction algorithm but also provides robustness to noise. However, in case of a high fraction of noise, CS fails to recover the sample efficiently, therefore, we propose to add another regularization term to our setup as explained below. More detail regarding CS methods and their variants, capabilities, properties and interesting applications can be found in [93–96].

3.4.3 Graph Total Variation

The graph total variation (GTV), denoted as $\|x\|_{gtv}$, in eq. (3.1), like the standard TV has two types 1) anisotropic and 2) isotropic. The former involves the sum of the gradients of nodes (entries in x) w.r.t $\mathcal{G}$ and the later involves the sum of the l_2 norms of the gradients at each node of $\mathcal{G}$. Throughout this chapter we use the former formulation as it has an intuitive interpretation, thus

$$\begin{aligned}\|x\|_{gtv} =& \|\nabla_{\mathcal{G}}(x)\|_1 = \sum_i \|\nabla_{\mathcal{G}} x_i\|_1 \\ =& \sum_i \sum_{j \in \mathcal{N}_i} \sqrt{W_{ij}} \|x_i - x_j\|_1, \end{aligned} \tag{3.2}$$

where the second sum runs over all the neighbors of i, denoted by $\mathcal{N}_i$. The above expression clearly states that GTV involves the minimization of the sum of the gradients of the signals on the nodes of the graphs. In our case, we assume that the elements of the vector x lie on the nodes of the graph $\mathcal{G}$ which are connected with the edges whose weights are W_{ij}. Thus, the minimization of the GTV would ensure that x_i and x_j possess similar values if W_{ij} is high and allow dissimilar values if W_{ij} is small or zero. As compared to the standard TV, the structure of the sample x is taken into account for the reconstruction purpose. It is a well known fact that l_1 norm promotes sparsity, so the GTV can also be viewed as a regularization which promotes sparse graph gradients. This directly corresponds to enforcing a smoothness of the signal x w.r.t graph $\mathcal{G}$.

The proposed method with GTV can be seen as a generalization of the compressed sensing and total variation based methods since standard TV commonly used with CS is just a special case of GTV. While, the standard TV minimizes the gradients of the signal x w.r.t its spatial neighbors only, the GTV does so in a region which is not restricted only to the neighbors of the elements in x. Thus, the standard TV can be viewed as a specific case of the GTV, where the graph $\mathcal{G}_{grid}$ is a grid graph. In a grid graph $\mathcal{G}_{grid}$ of a sample x, the pixels are only connected to its spatial neighbors (upper, lower, left and right) via unity weights.

Using eq. (3.2) in eq. (3.1), our proposed model can be written as:

$$\min_c \|Ax - b\|_2^2 + \lambda \|\Phi^*(x)\|_1 + \gamma \|\nabla_{\mathcal{G}} x\|_1 \tag{3.3}$$

Now using $x = \Phi(c)$ we get

$$\min_c \|A\Phi(c) - b\|_2^2 + \lambda \|c\|_1 + \gamma \|\nabla_{\mathcal{G}}(\Phi c)\|_1 \tag{3.4}$$

From the above equation it is obvious that our proposed model promotes simultaneous sparsity of the transform coefficients c in the Wavelet domain Φ and the sparsity of the gradients in the graph domain. Hence, the model can be viewed as promoting a doubly sparse structure in wavelet and graph domains respectively.

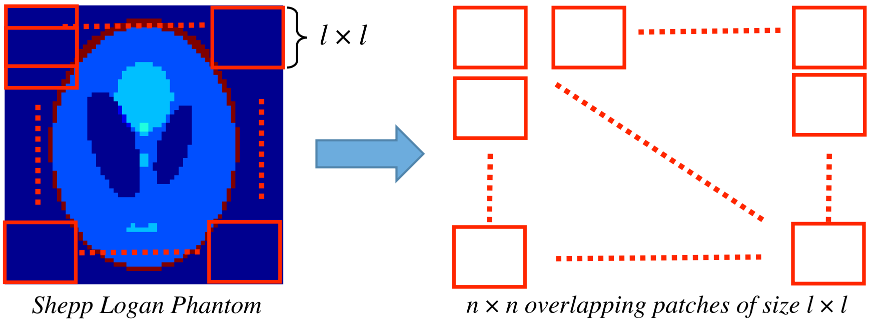


Figure 3.1: For the construction of patch graph, $x_{fbp} \in \mathbb{R}^{n \times n}$ is divide into n^2 overlapping patches of size $l \times l$ each.

3.4.4 Graph Construction for Total Variation

An important step for our method is to construct a graph $\mathcal{G}$ for TV regularization. In contrast to standard TV, which can be directly used as a prior for regularization, GTV needs a graph $\mathcal{G}$ to start with. Ideally, $\mathcal{G}$ should be representative of the reconstructed sample x, however, this is unknown before the reconstruction. To cater this problem, we propose to construct $\mathcal{G}$ from an initial naive estimate of the sample x_{fbp} using filtered back projection (FBP) method.

We propose to construct a graph between the patches of x_{fbp} instead of pixels. As the sample to be reconstructed is an image, and the graph is being constructed from the noisy x_{fbp}, to obtain robustness to noise, it makes more sense to construct the graph from the patches rather than the pixels of x_{fbp}. In the first step $x_{fbp} \in \mathbb{R}^{n \times n}$ is divided into n^2 overlapping patches. Let s_i be the patch of size $l \times l$ centered at the i^{th} pixel of x_{fbp} and assume that all patches are vectorized, $i.e$, $s_i \in \Re^{l^2}$ (Figure 3.1). In the second step the search for the closest neighbors for all vectorized patches is performed using the Euclidean distance metric. Each s_i is connected to its $\mathcal{K} = 10$ nearest neighbors s_j, resulting in $|\mathcal{E}|$ number of connections. In the third step the graph weight matrix W is computed using the Gaussian kernel weighting scheme (eq. (2.1)), for which the parameter σ is set experimentally as the average distance of the connected samples. Finally, the combinatorial Laplacian is computed.

Note that the computation of the weight matrix W for graph $\mathcal{G}$ costs $\mathcal{O}(n^4)$. For small n^2, we can use the above strategy directly. Although, the computation of W is expensive, it should be noted that with sufficiently small n^2, the graph can still be computed in the order of a few seconds. For big or high dimensional samples, $i.e$, large n^2, we can use a similar strategy but the computations can be made efficient ($\mathcal{O}(n^2 \log n^2)$) using the FLANN library (Fast Library for Approximate Nearest Neighbors searches in high dimensional spaces) [91]. However, the quality of the graphs constructed using this strategy is slightly lower due to the approximate nearest neighbor search method.

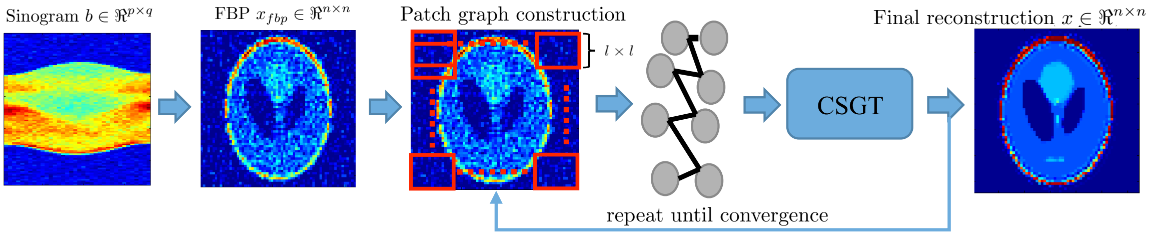

Sinogram $b \in \Re^{p \times q}$
FBP $x_{fbp} \in \Re^{n \times n}$
Patch graph construction
$l \times l$
CSGT
Final reconstruction $x \in \Re^{n \times n}$
repeat until convergence

Figure 3.2: The complete methodology for AGTV. The input sinogram / projections $b \in \mathbb{R}^{p \times q}$ is first used to obtained a filtered back projection (FBP) $x_{fbp} \in \mathbb{R}^{n \times n}$. It is then used to construct the initial patch graph $\mathcal{G}$ to be used by the CSGT method. The output of CSGT is used to refine / reconstruct the graph and this process is repeated until convergence.

3.4.5 Adaptive Graph Total Variation Regularization

The above description refers only to the non-adaptive part, where the graph $\mathcal{G}$ is fixed. It is important to point out that the initial estimate of the graph $\mathcal{G}$, obtained via the filtered back projection x_{fbp} is not very faithful to the final solution x. As x is being refined in every iteration, it is natural to update the graph $\mathcal{G}$ as well in every iteration. This simultaneous update of the graph $\mathcal{G}$ corresponds to the adaptive part of the proposed algorithm and its significance will be explained in detail in Section 3.6 of the chapter.

3.5 Optimization Solution

We make use of proximal splitting methods to solve problem (3.3). The specialty of such methods is that they can resolve a tedious complex problem into smaller and relatively trivial subproblems which are solved using proximal operators. The proximal operator of a function λh is defined as follows:

$$\mathrm{prox}_{\lambda h}(y) = \underset{x}{\mathrm{argmin}} \frac{1}{2}\|x - y\|_2^2 + \lambda h(x).$$

More detailed information about such methods can be found in [57, 97].

3.5.1 Forward-Backward based Primal Dual Solver

We cast our problem in the form and use the Forward-backward based primal dual method to solve it.

$$\underset{x}{\mathrm{argmin}}\, f(x) + g(Ax) + h(B(x)). \tag{3.5}$$

The first term of (3.1), $f : \mathbb{R}^{n^2} \to \mathbb{R}$ is a convex differentiable function defined as $f(x) = \|Ax - b\|_2^2$. This function has a β-Lipschitz continuous gradient

$$\nabla_f(x) = 2A^\top(Ax - b).$$

Note that $\beta = 2\|A\|_2$ where $\|A\|_2$ is the spectral norm (or maximum eigenvalue) of A. The constant β has important implications in deciding the time step in the iterative optimization methods.

The proximal operator of the second function $g = \lambda\|\Phi^*(x)\|_1$ (in eq. 3.3) is the ℓ_1 soft-thresholding of the wavelet coefficients given by the elementwise operations.

$$\text{prox}_{\tau_1 g}(x) = \text{sgn}(x) \circ \max(|x| - \tau_1\lambda, 0), \tag{3.6}$$

The third term in eq. (3.3) $h : \mathbb{R}^{|\mathcal{E}|n} \to \mathbb{R}$, where $|\mathcal{E}|$ denotes the cardinality of $\mathcal{E}$ the set of edges in $\mathcal{G}$, is a convex function defined as $h(D) = \gamma\|D\|_1$. The proximal operator is:

$$\text{prox}_{\tau_2 h}(D) = \text{sgn}(D) \circ \max(|D| - \tau_2\gamma, 0), \tag{3.7}$$

where $\circ$ denotes the Hadamard product and $D = \nabla_{\mathcal{G}} x$. Further details regarding forward backward solvers have been reviewed in [57, 98–100].

3.5.2 Algorithm

Using these tools, we can use the forward backward based primal dual approach presented in [97] to define Algorithm 2 where τ_1, τ_2, τ_3 are convergence parameters ϵ the stopping tolerance and I, J the maximum number of iterations. δ is a very small number to avoid a possible division by 0. Since we use Unlocbox [58] for solving the optimization algorithm, the convergence parameters τ_1, τ_2, τ_3 are set automatically according to the specified β. U_j corresponds to the primal and V_j to the dual variable in Algorithm 2.

More specifically, Algorithm 2 is based on a forward-backward approach [57]. It combines a gradient descent step (forward step) with a computation step involving a proximity operator (step 1a in Algorithm 2). Note that the gradient in this step is w.r.t. the differentiable function f, to which the result of the application of adjoint operator of g, *i.e*, $\nabla^*_{\mathcal{G}}$ is added. Then, the proximity step corresponds to the application of the proximal operator of h, which is an element wise soft-thresholding, on this result. This computation corresponds to a kind of subgradient step performed in an implicit (or backward) manner [57]. A deeper justification of this terminology is provided by the theory of monotone operators [100].

3.5.3 Computational Complexity

Complexity of Graph Construction

As mentioned earlier we use the Fast Approximate Neartest Neighbors search algorithm (FLANN) [91]. The computational complexity of the FLANN algorithm for n^2 patches of size l^2 each and fixed $\mathcal{K}$ is $\mathcal{O}(n^2 \log(n^2))$. Note that l^2 and $\mathcal{K}$ do not appear in the complexity because they are constants. Furthermore, n^2 is the size of the sample under consideration so the computational complexity is much lower than NLTV [52] based methods.

Algorithm 2 Forward-backward primal dual for AGTV

```
$x_0 = x_{fbp}$
1. INPUT: $U_0 = x_0$, $V_0 = \nabla_{\mathcal{G}} x_0$, $\epsilon > 0$
for $i = 0, \dots I - 1$ do
    for $j = 0, \dots J - 1$ do
        a. $P_j = \Phi(\mathrm{prox}_{\tau_1 g}\left(\Phi^*(U_j) - \tau_1 \Phi^*\left(\nabla_f(U_j) + \nabla_{\mathcal{G}}^* V_j\right)\right)$
        b. $T_j = V_j + \tau_2 \nabla_{\mathcal{G}}(2P_j - U_j)$
        c. $Q_j = T_j - \tau_2 \,\mathrm{prox}_{\frac{1}{\tau_2} h}\left(\frac{1}{\tau_2} T_j\right)$
        d. $(U_{j+1}, V_{j+1}) = (U_j, V_j) + \tau_3((P_j, Q_j) - (U_j, V_j))$
        if $\frac{\|U_{j+1} - U_j\|_F^2}{\|U_j\|_F^2 + \delta} < \epsilon$ and $\frac{\|V_{j+1} - V_j\|_F^2}{\|V_j\|_F^2 + \delta} < \epsilon$ then
            BREAK
        end if
    end for
    2. $x_i = U_{j+1}$
    3. Construct patch graph $\mathcal{G}$ from $x_i$
    if $\frac{\|x_i - x_{i-1}\|_F^2}{\|x_i\|_F^2 + \delta} < \epsilon$ then
        BREAK
    end if
end for
OUTPUT: $x_i$
```

Algorithm Complexity

Let J denote the number of iterations for the algorithm (the for loop in Algorithm 2) to converge, and I the number of outer iterations (step 4 of Algorithm 2), then the computational cost of our algorithm is $\mathcal{O}((J|\mathcal{E}|I)$, where $|\mathcal{E}|$ denotes the number of non-zeros edges in the graph $\mathcal{G}$. For a $\mathcal{K}$-nearest neighbors graph $|\mathcal{E}| \approx \mathcal{K}n^2$ so the computational complexity of our algorithm is linear in the size of the data sample n^2, *i.e* $\mathcal{O}((J\mathcal{K}n^2I)$.

Overall Complexity

The complexity of our algorithm is $\mathcal{O}((J\mathcal{K}n^2I)$ and the graph $\mathcal{G}$ is $\mathcal{O}(n^2\log(n^2))$. The graph $\mathcal{G}$ needs to be updated once in every outer iteration of the algorithm I, thus the overall complexity of the proposed AGTV method is $\mathcal{O}(IJ\mathcal{K}n^2 + In^2\log(n))$.[4] The complexity term does not depend on the image size. The iterations in Algorithm 2 are independent of image size. However, it takes longer for the optimization problem to converge for large size images, as will be shown in the experiments.

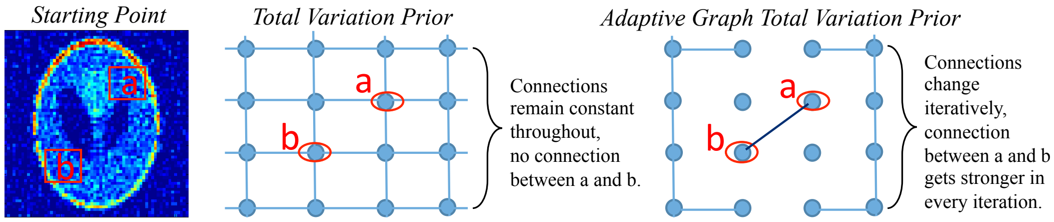

Starting Point
a
b
Total Variation Prior
a
b
Connections remain constant throughout, no connection between a and b.
Adaptive Graph Total Variation Prior
a
b
Connections change iteratively, connection between a and b gets stronger in every iteration.

Figure 3.3: A comparison of the Total Variation (TV) and Adaptive Graph Total Variation (AGTV) priors for the methods CSGT and AGTV. The TV prior does not connect patches 'a' and 'b' which possess structural similarity, whereas the GTV prior connects them because the $\mathcal{K}$-nearest neighbor graph is not restricted to spatial neighbors only. Furthermore, this connection keeps getting stronger due to iterative removal of noise and graph updates in every iteration.

3.6 Working Explanation of AGTV

We present a simple example to motivate the use of AGTV rather than simple CSGT and CSTV[5]. Clearly, the compressed sensing part of all these methods is responsible for retrieving the sample x from the projections b. Thus, our comparison study is focused on the two regularizers, *i.e*, Adaptive Graph Total Variation (AGTV) and Total Variation (TV). Our two step exposition is described below:

1. CSGT is better than CSTV.
2. Adaptive Graph Total Variation (AGTV) is better than CSGT.

Consider the example of a Shepp-Logan Phantom as shown in top leftmost plot of Fig. 3.4. The goal is to recover this phantom from its noisy projections so that the recovered sample is faithful to its original clean version. The CSTV method requires a total variation prior to recover the sample while the CSGT method requires a graph total variation prior for the recovery. Both methods need an initial estimate for the construction of this prior, therefore, for the ease of demonstration we use the filtered back projection (FBP) as an initial estimate of the sample. Recall that our proposed method decomposes the FBP into $n \times n$ patches of size $l \times l$ each. Let (i, j) denote the (horizontal, vertical) position of the center of each patch then:

- For the total variation, each patch $s_{i,j}$ is connected to its spatial neighbors only, *i.e*, $s_{i+1,j}, s_{i-1,j}, s_{i,j+1}, s_{i,j-1}$, as shown in Fig. 3.3. These connections are fixed throughout the algorithm.
- For the graph total variation, each patch $s_{i,j}$ is only connected to the patches which are among the $\mathcal{K}$ nearest neighbors. Note that unlike TV the connected patches can be spatially far from each other.

[4] The square term in the log complexity may be further simplified: $\mathcal{O}(n^2 \log(n^2)) = \mathcal{O}(2n^2 \log(n)) = \mathcal{O}(n^2 \log(n))$.

[5] The acronyms can get confusing here. AGTV is the proposed method which is a combination of CS and adaptive Graph TV. CSTV is the classical method of using CS and local TV. CSGT is CS and Graph TV (*i.e.*, similar to AGTV but similarity matrix is not updated during every iteration).

Now let us take the example of two patches ‘a’ and ‘b’ as labeled in the FBP of Fig. 3.3. Comparing with the clean phantom in Fig. 3.4 it is obvious that these patches should possess the same texture at the end of the reconstruction algorithm. Therefore, an intelligent regularizer should take into account the inherent similarity between these patches. To explain the difference between the TV and GTV priors we use a point model as shown in Fig. 3.3, where each point corresponds to a patch in the FBP. Since ‘a’ and ‘b’ are not spatially co-located, the total variation prior does not establish any connection between these patches. Thus, TV fails to exploit the similarity between these patches throughout the algorithm. This leads to slightly different textures for the two patches, as shown in the 3rd row of Fig. 3.4.

Now consider the case of GTV. Even though the intial estimate of graph $\mathcal{G}$ is obtained from the noisy estimate of sample, *i.e*, the FBP, patches ‘a’ and ‘b’ still possess enough structural resemblance to be connected together by an edge (even if it is weak) in the graph. Now, if the graph is kept fixed which is the case of CSGT, one still obtains a better result as compared to CSTV, as shown in the 5th row of Fig. 3.4. This is due to the fact that the important connections are established by the graph $\mathcal{G}$ and similarity of patches is not restricted to spatially co-located patches only. This is also obvious from the intensity profile analysis in the 6th row of Fig. 3.4. Finally, we discuss the case of AGTV, where the graph $\mathcal{G}$ is updated in every iteration of the algorithm. Obviously, every iteration of the algorithm leads to a cleaner sample and updating the graph $\mathcal{G}$ is only going to make the connection between the patches ‘a’ and ‘b’ stronger. This leads to significantly better result than CSTV and CSGT as shown in Fig. 3.4. Note that the patches ‘a’ and ‘b’ possess almost the same structure at the end of AGTV.

3.7 Experimental Results

General Experimental Setup

Experiments were performed using two open source toolboxes, GSPBox [101] for the graph construction and UNLocBox for the convex optimization part [58]. These toolboxes provide fast and efficient general purpose algorithms, with an automatic tuning of many implicit non-model parameters, such as step size for iterative optimization algorithms. Additional code was written in MATLAB to enable AGTV with these packages. To test the performance of our AGTV method, we perform reconstructions for many different types of phantoms from different number of projections with varying levels of Poisson noise. Throughout this section, we report the reconstruction results of various phantoms in terms of the ℓ_2 reconstruction error. We compare the performance of AGTV with many state-of-the-art iterative and convex optimization based algorithms, which include FBP, ART (Kaczmarz), SIRT (Cimmino), CS, CSTV and CSGT. Simulation setup pertaining to more realistic Cryo-ET simulations which incorporate the effects of the missing wedge and simulated shift between the projections was according to Experiment 2.3 presented in Chapter 2. Each of these methods has its own model parameters, which need to be set or tuned in an appropriate manner.

Experiment 3.1: Effectiveness of AGTV when compared to algebraic and state-of-the-art SEIR methods.

Aim

To demonstrate the effectiveness of reconstructing a noisy Shepp-Logan phantom from limited data using AGTV when compared to algebraic methods (ART, SIRT) and SEIR methods (CS,CSTV,CSGT).

Method

To explain the performance of our model in detail we reconstructed a 64×64 Shepp-Logan [26] phantom from 36 erroneous projections. A sinogram S was built by projecting the phantom using Radon transform and 36 equally spaced projections were collected from 0 to 180 degrees. The projections were then corrupted with 10% Poission noise. ART (Kaczmarz) and SIRT (Cimmino) were performed using FBP as *a priori*. The stopping criteria for ART and SIRT was set to 100 iterations and the relaxation parameter (η) was tuned to achieve the best result. For the CS method, the reconstruction was performed for uniformly spaced values of λ in the range $(0, 10)$ and the best λ was selected based on the minimum ℓ_2 reconstruction for the phantoms. For CSTV, the reconstruction was performed for sparsity parameter $\lambda \in (0, 1)$ and TV parameter $\gamma \in (0, 10)$ and the parameters for the best result were selected. The ranges were defined linearly following a more exhaustive logarithmic range search. For the graph based reconstruction (CSGT, AGTV) a graph prior $\mathcal{G}$ was generated by dividing the result from FBP into patches as explained in Section 3.4.4. For example, for a Shepp-Logan phantom of size 64×64, the graph was constructed by dividing it into $64 \times 64 = 4096$ overlapping patches of size 3×3. A graph $\mathcal{G}$ was constructed between the 4096 patches with 15 nearest neighbors ($\mathcal{K} = 15$) and σ for the weight matrix was set to the average distance of the 15-nearest neighbors. The number of nearest neighbors is usually set via hit and trial. More complex methods for adjusting the value of $\mathcal{K}$ is an open research problem. Setting K too large makes the graph fully connected which leads to operations being computational complex. Setting $\mathcal{K}$ too small may overlook similarities between different regions of the image. Experiment 3.5 discusses this in more detail.

Randomized Kaczmarz used for comparison in this experiment is a variant of Kaczmarz method whereby the rows associated ith $Ax = b$ are accessed at random [102]. For Algorithm 2, we set $I = J = 50$ and the convergence parameters τ_1, τ_2, τ_3 were set automatically by UNLocBox. It is worth mentioning here that our GTV based adaptive graph regularization is a faster method of implementing NLTV by using $\mathcal{K}$-nearest neighbors graph. Thus the CSGT and NLTV based regularization are equivalent in performance. Therefore, we did not include comparisons with the NLTV based method.

Conclusion

Figure 3.4 provides a detailed comparison of the reconstruction of Shepp-Logan phantom via various algorithms along with the intensity profiles plotted underneath each of the reconstructions. It can be clearly seen that AGTV performs better than CSGT and

CSTV. It is possible to appreciate this visually as the phantom obtained via AGTV is very similar to the original phantom except for the edges. The edges may be corrected by tuning the parameters better. Furthermore, a comparison of the intensity profiles of the two phantoms also reveals the same fact. The next best result is obtained by CSGT. Algorithmically, the only difference between CSGT and AGTV is the regular graph update step in the latter, which tends to make the final reconstruction more faithful to the original phantom. CSTV also obtains a reasonable reconstruction, though worse than AGTV. CS alone however, has a poor performance. This is not surprising, as for tomographic applications, CS has been mostly used in combination with TV, as it alone does not cater for the sparsity required. The claim that ACSGT is better than other methods is further substantiated in Experiment 3.2.

Experiment 3.2: Effectiveness of AGTV with a 128×128 Torso Phantom

Aim

To demonstrate the effectiveness of reconstructing a noisy 128×128 Torso phantom from limited data using AGTV with reduced number of nearest neighbors $\mathcal{K}$ and fewer iterations when compared and SEIR methods (CS,CSTV,CSGT).

Method

A 128×128 Torso phantom was simulated to generate from 36 projections. A sinogram S was built by projecting the phantom using Radon transform and 36 equally spaced projections were collected from 0 to 180 degrees. The projections were then corrupted with 5% Poission and Gaussian noise. A graph $\mathcal{G}$ was constructed between the 4096 patches with 15 nearest neighbors ($\mathcal{K} = 10$) and σ for the weight matrix was set to the average distance of the 10-nearest neighbors. For Algorithm 2, we set $I = J = 25$ and the convergence parameters τ_1, τ_2, τ_3 were set automatically by UNLocBox. The rest of the simulation setup was exactly the same as Experiment 3.1.

Conclusion

Fig. 3.5 shows the result of reconstructing the standard Torso phantom using various SER methods. It can be clearly seen that AGTV preserves the edges much more as compared to CSTV and CSGT. This makes sense because one of non-local connections between various regions of the image as explained in 3.6. In contrast to experiment 3.1 this experiment uses a relatively larger phantom which comes at the expense of greater computational run-time, *i.e.*, 3874sec (≈64 min) as opposed to 53.7 sec (<1 min) on an Intel Xeion Machine (E5450 12M Cache, 3.00 GHz, 128GB RAM). The difference is stark, this is due to the computational complexity associated with graph construction during every iteration. Further examining Algorithm 2 also reveals that there are a substantial number of multiplications associated with various operations which drive the computational time higher. The increase in computational run time is despite the fact that the nearest neighbor search is reduced to 10 patches and the number of iterations were reduced ($I = J = 25$). That said, the result with the proposed method is still of substantial quality when reconstructing from just 36 projections. If

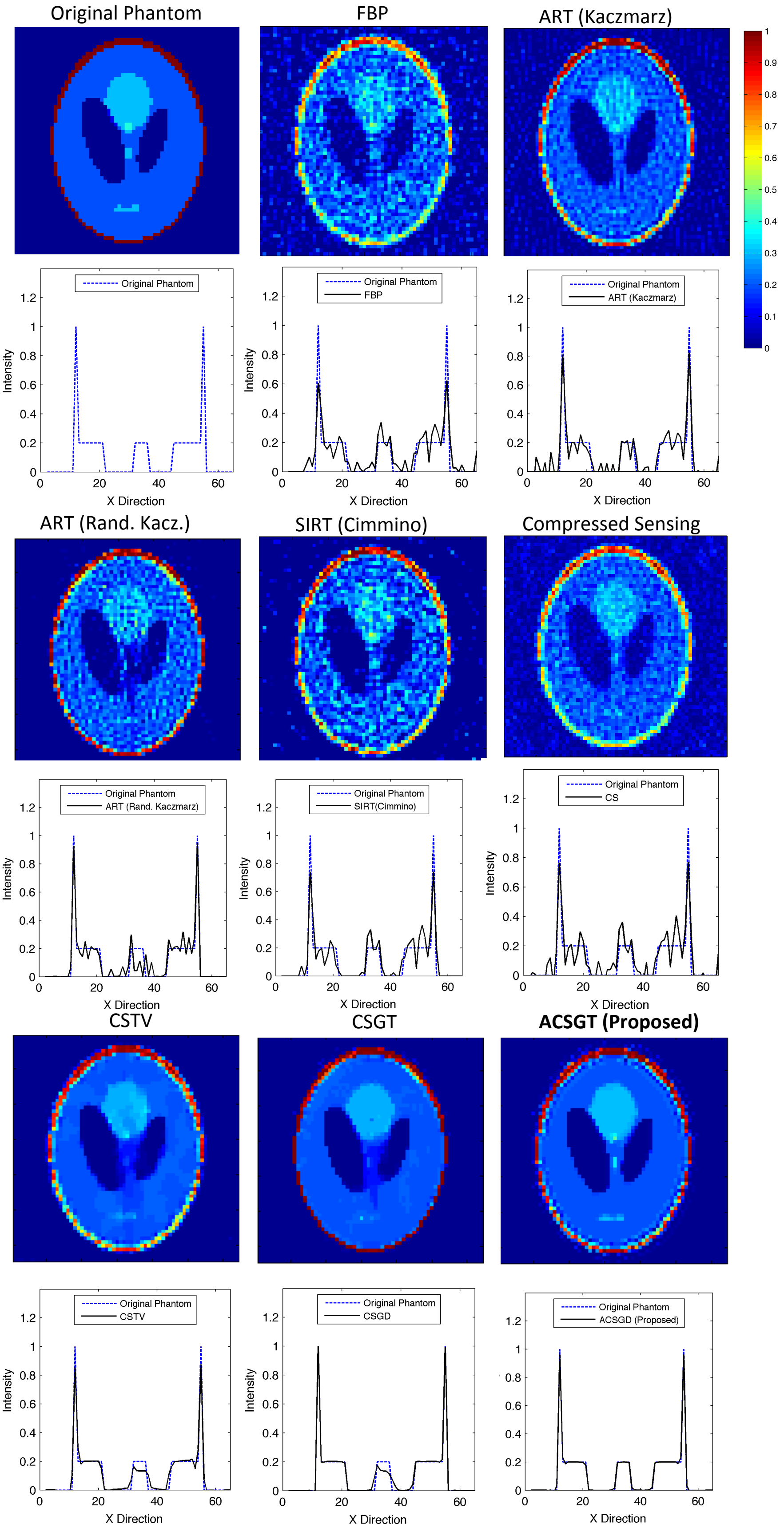


Figure 3.4: Comparative analysis of reconstructions and intensity profiles of reconstructing a Shepp-Logan phantom using various reconstruction methods.

the similarity matrix is not updated every iteration CSTV often behaves very similar to CSGT, especially at lower values of $\mathcal{K}$, this effect can be seen from the results.

Experiment 3.3: Effectiveness of AGTV at Various Number of Projections

Aim

To study the ability of AGTV to recover a phantom from less number of projections. The basic premise of CS-type methods is their ability to recover data using very few measurements. Being a CS-type method which uses non-local processing and promotes sparsity in wavelet and graph gradient domains, AGTV should be able to recover missing data. In order for AGTV to demonstrate superiority over CSTV and CSGT (or NLTV) it should be able to recover data from few measurements while keeping a lower error measure as compared to competitive SER methods.

Method

A 128×128 Shepp Logan phantom was simulated with varying number equally spaced projections between 9 and 180 and 6 different sinograms were generated. The projections were corrupted with 5%, 10% and 20% Poisson Noise, finally generating 18 different sinogram. Each of these 18 sinograms was then reconstructed using FBP (Linearly Interpolated, Ram-Lak Filtered), ART (Kaczmarz), SIRT (Cimmino), CS, CSTV, CSGT (NLTV), ACSGT. Error Norms were calculated against the original phantom for all 126 reconstructions were calculated using $\|x - x^*\|_2$, where x is the result from a specific method and x^* is the original phantom. The two hyper-parameters were tuned to give the best result for each reconstruction (*i.e.*, almost 5-10 reconstructions for approximately tuning each individual reconstruction)[6].

Conclusion

A graphical comparison of the 126 reconstruction mentioned above is provided in Fig. 3.6. It is clear that AGTV method generally performs better as compared to many other state-of-the-art methods and follows a similar trend with respect to the number of projections. It is also interesting to note that the performance of AGTV saturates after 90 projections for each of the three cases, i.e, the reconstruction error does not improve if the number of projections are increased. Furthermore, for each of the three noise cases one can observe that the drop in the reconstruction error from 50 to 90 projections is not significant. Although, the same observation can be made about CSGT, the error is a always higher than AGTV. All the other methods, perform far worse than AGTV. These graphs clearly, lead to the conclusion that AGTV is a step towards getting very fine reconstructions from a very small number of projections, via a scalable method. That said, these reconstruction are with equally spaced missing data. The effects of the method on continuous missing data will be investigated in the next experiment.

[6] Fun fact: This was the longest running image reconstruction simulation in the entire thesis!

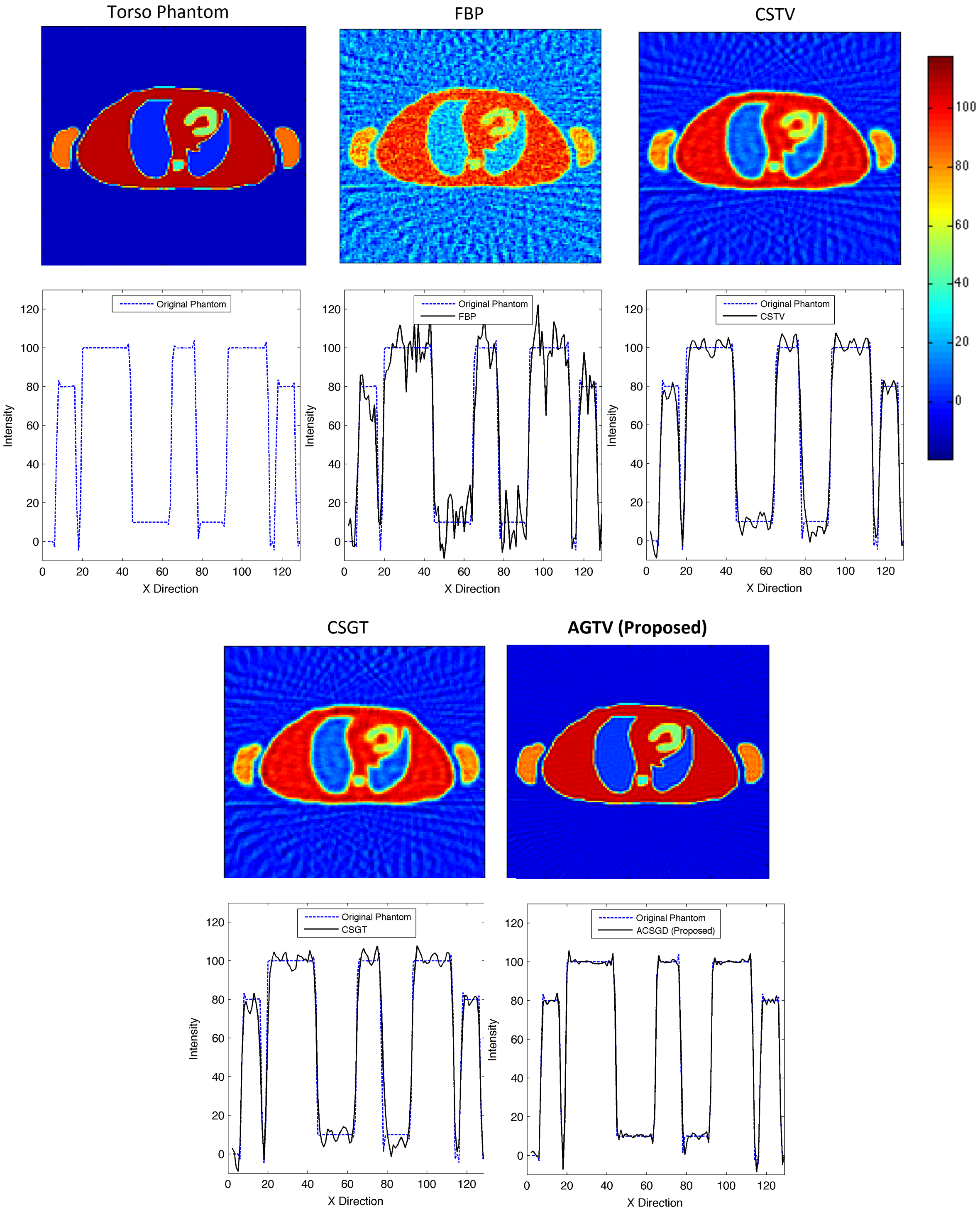
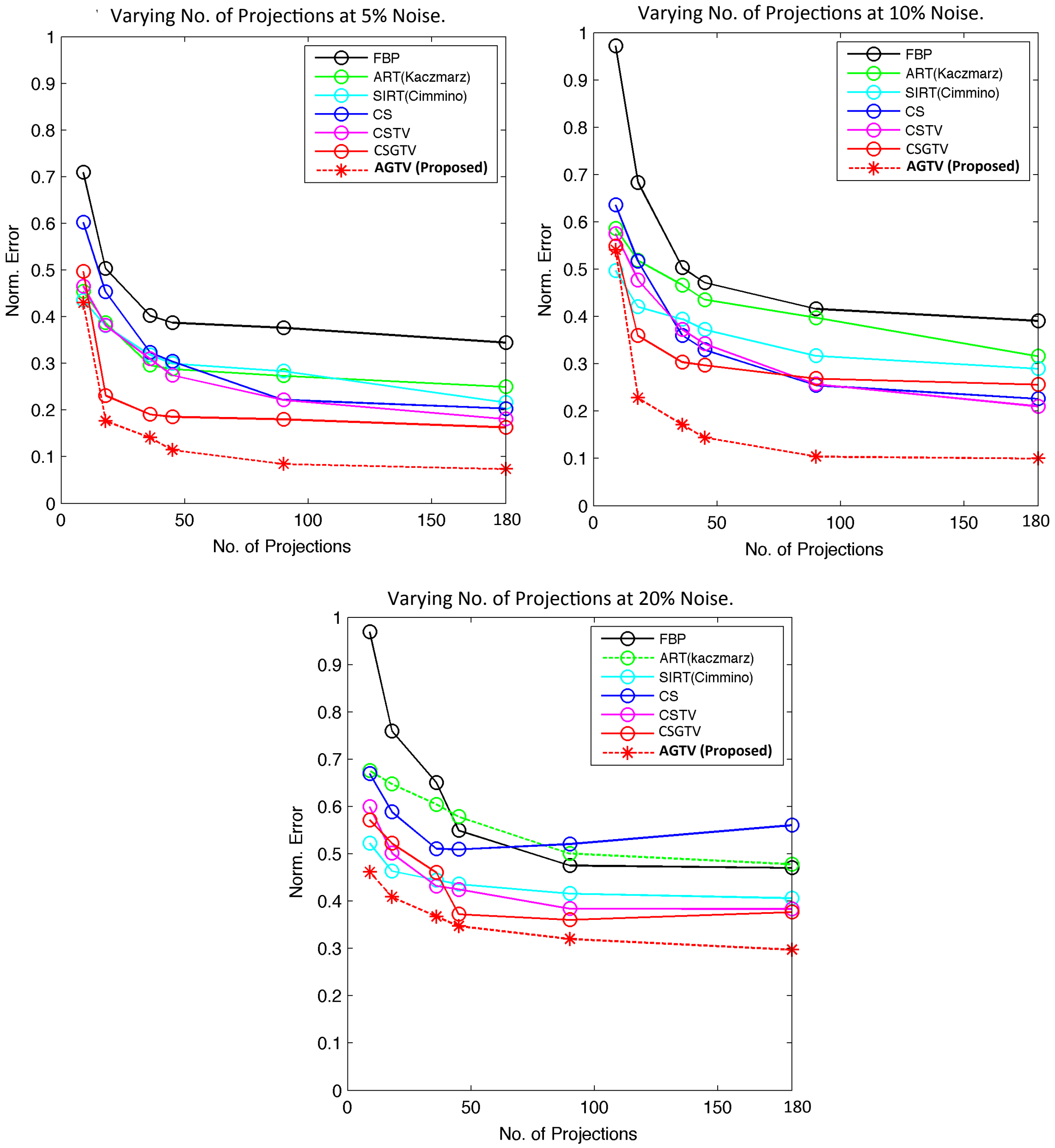


Figure 3.5: Comparative analysis of reconstructing a Torso phantom using various reconstruction methods.



Varying No. of Projections at 5% Noise.

Varying No. of Projections at 10% Noise.

Varying No. of Projections at 20% Noise.

Figure 3.6: Comparative analysis of reconstructing a Shepp-Logan phantom using various reconstruction methods at 5% and 10% and 20% Poisson noise. FBP (Linearly interpolated, Cropped Ram-Lak filter); ART (Kaczmarz/Randomized Kaczmarz, Relaxation Parameter (η) = 0.25, Prior: FBP, Stopping Criteria = 100 iterations); SIRT (Cimmino/SART, (η) = 0.25, Prior: FBP, Stopping Criteria = 100 iterations); CS (500 Iterations, Prior: FBP); CSTV ($\lambda = 0.5$, $\gamma = 0.1$, Prior: FBP, Stopping Criteria = 100 iterations); CSGT ($\lambda = 0.5$, $\gamma = 0.2$, Prior: Patch Graph from FBP, Stopping Criteria = 100 iterations); AGTV ($\lambda = 0.5$, $\gamma = 1$, Prior: Patch Graph from FBP updated every iteration, I and J in Algorithm 2 set to 30).

Experiment 3.4: AGTV Reconstructions and the Missing Wedge

Aim

To investigate the ability of AGTV to reconstruct components of the missing wedge and to determine the extent of high frequency components lost during signal recovery while employing AGTV. We have seen that being a SEIR method AGTV has the ability to recover components of the missing data. This section will investigate how does the method behave in a more stark reduced data situation.

Method

The simulation setup for emulating 2D Cryo-ET phantom simulations accurately presented in Experiment 2.3 was used. The model has the ability to accurately simulate the missing wedge and problems associated with the alignment of data. A 128×128 Shepp-Logan phantom was simulated to produce 71 equally spaced projections with a $-70^\circ \leq \theta \leq +70^\circ$ missing wedge. The resulting projections were corrupted with 5% Noise, the rest of the simulation setup is according to Experiment 3.3. Besides the intensity profiles, radially averaged power spectrum explained in Chapter 2 and Fourier ring correlation were used for frequency and resolution preservation analysis. Fourier ring correlation is the 2D version of Fourier shell correlation, a metric commonly used in structural biology to determine the resolution of 3D structures. In this case it will simply be used to correlate the frequency spectrum of the original Shepp-Logan phantom with various methods. All parameters were individually tuned to achieve best results.

Conclusion

Fig. 3.7 shows a comparative analysis of reconstructions and intensity profiles in both directions of the Shepp-Logan phantom using various reconstruction methods at 5% noise. It can clearly be seen that AGTV behaves better than all other methods, although it does not fully reconstruct the missing wedge it is able to partly reconstruct it. This is due to the ability of SEIR-based methods to reconstruct from missing and noisy data which is enhanced by the fact that AGTV promotes sparsity in dual graph gradient and wavelet domains. Fig. 3.8 shows the Fourier ring correlation of individual methods with the original Shepp-Logan phantom, this clearly shows that there is very little loss in resolution in the case of AGTV[7]. The radially averaged power spectrum follows a similar trend whereby AGTV experiences minimum high resolution loss when compared to CSGT and other methods. Finally Fig. 3.10 shows the 2D FFT magnitude spectrum and the extent to which the missing wedge can be recovered. The recovery is attributed to the fact that AGTV uses CS-type reconstruction where by the reconstruction problem is seen as a compressed image in some unknown sparse bases. AGTV is able to harness this idea and recover part of the missing data. It is worth mentioning that although the Shepp-Logan phantom could be recovered the missing wedge problem for real data is much more complex since it is much harder to cater to the sparsity necessary to enable such a recovery specifically for biological experiments.

[7] An anonymous reviewer deserves to be acknowledged for suggesting the use of FRC.

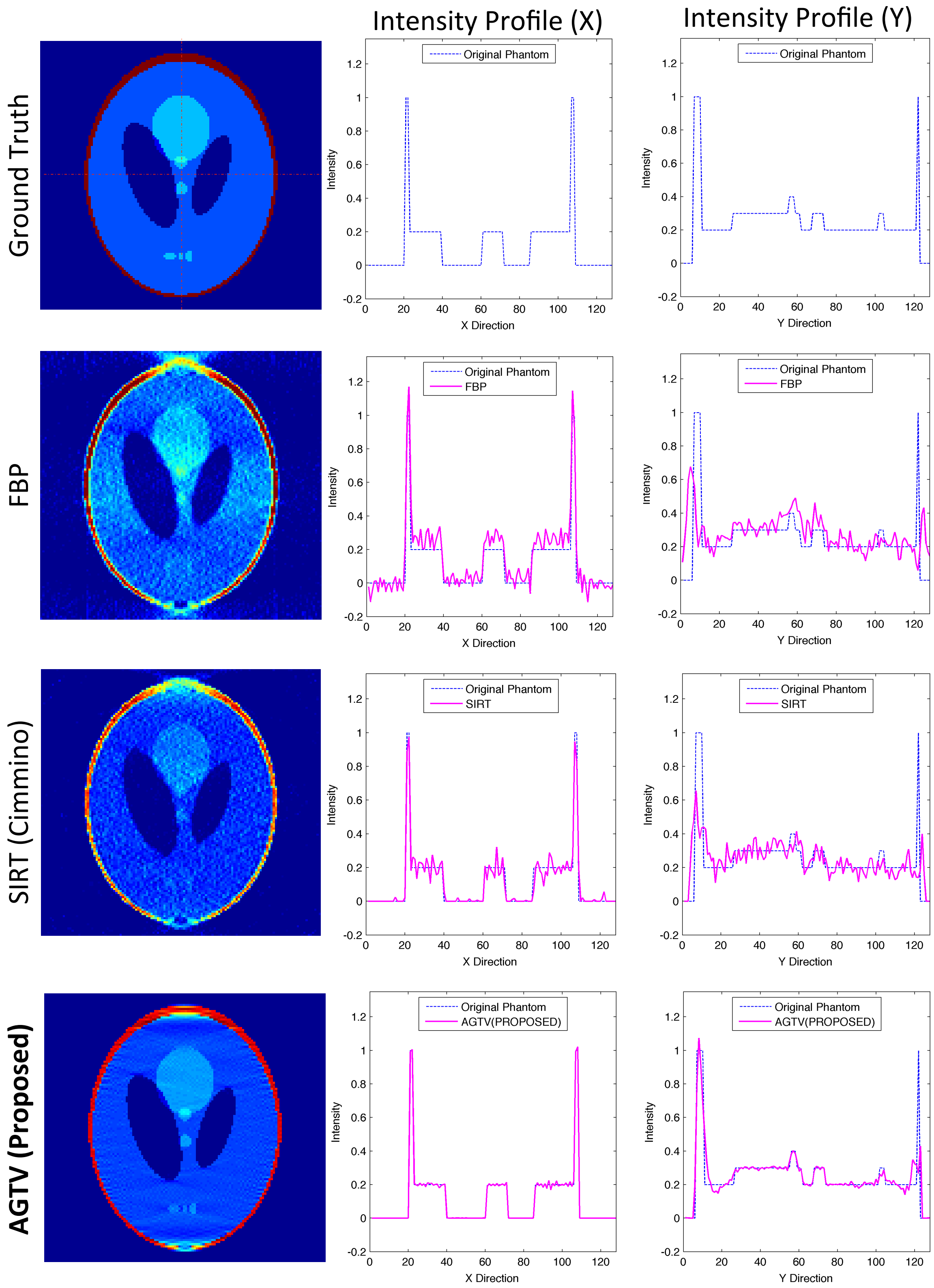


Figure 3.7: Comparative analysis of reconstructions and intensity profiles in both directions of a Shepp-Logan phantom using various reconstruction methods at 5% noise. It can clearly be seen that AGTV behaves better than all other methods, although it does not fully reconstruct the missing wedge it is able to partly reconstruct it.

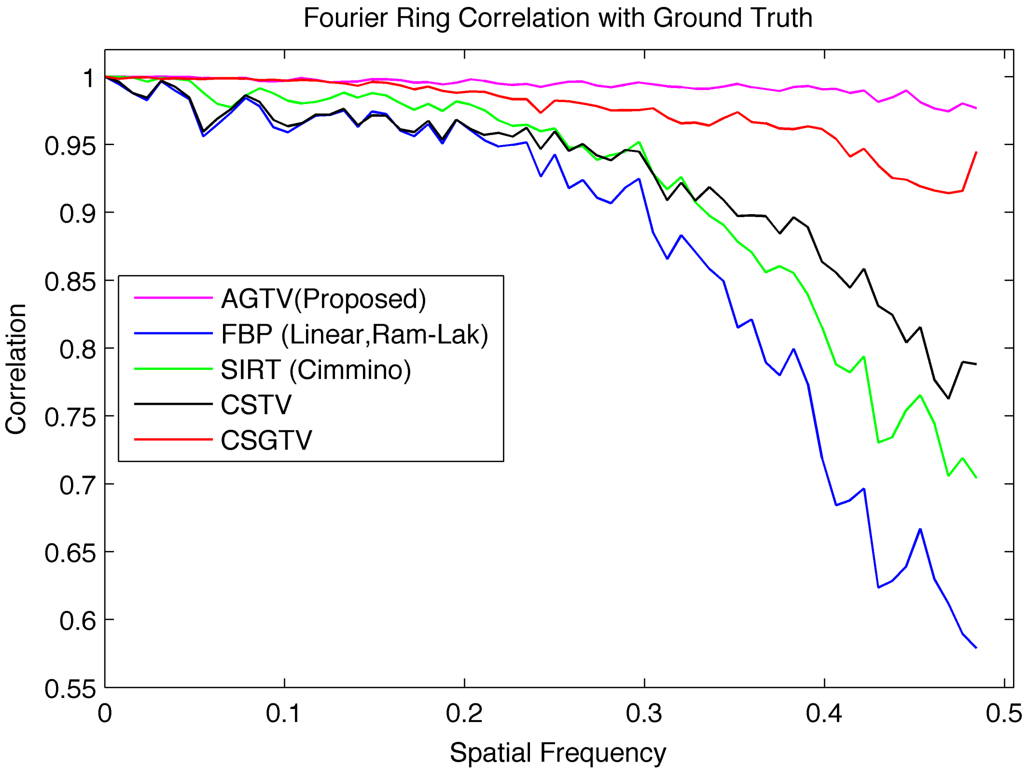


Figure 3.8: Fourier Ring Correlation of the original Shepp-Logan phantom with reconstructions from various reconstruction methods. Clearly there is hardly any loss of resolution in the case of AGTV, mild loss is expected since the method cannot perfectly reconstruct the missing wedge and all noise cannot be removed.

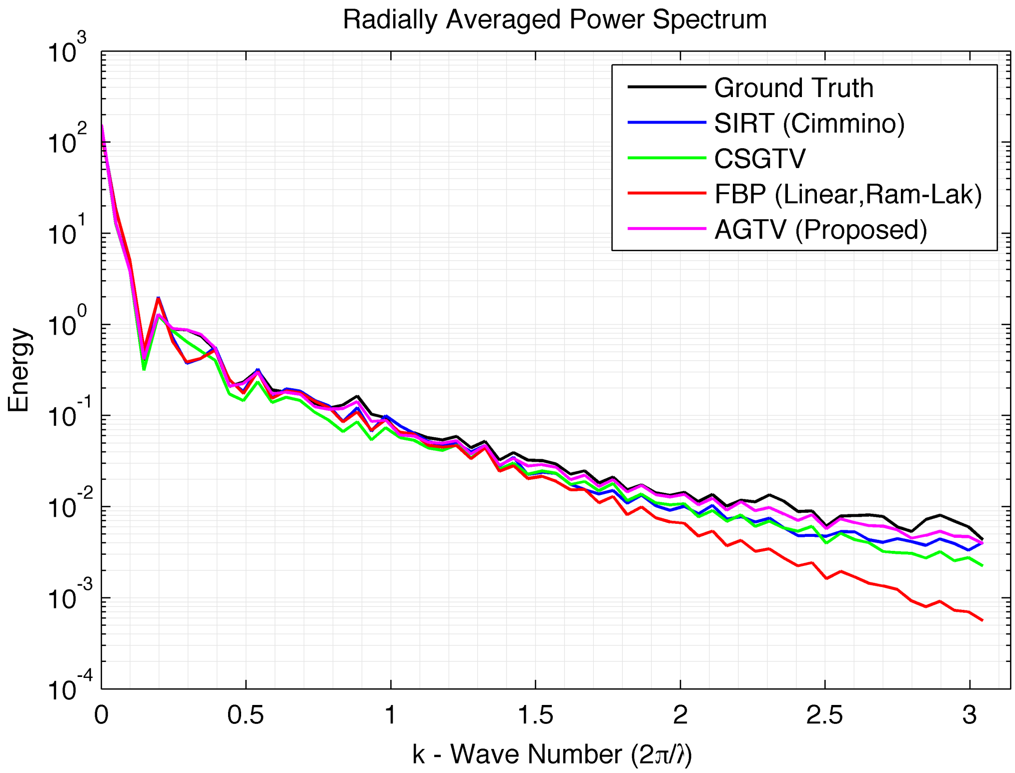


Figure 3.9: Comparative analysis of the radially averaged power spectrums of the Shepp Logan phantom reconstructed employing various reconstruction methods. It can be seen that AGTV is the closest to the spectrum of the Shepp-Logan ground truth.

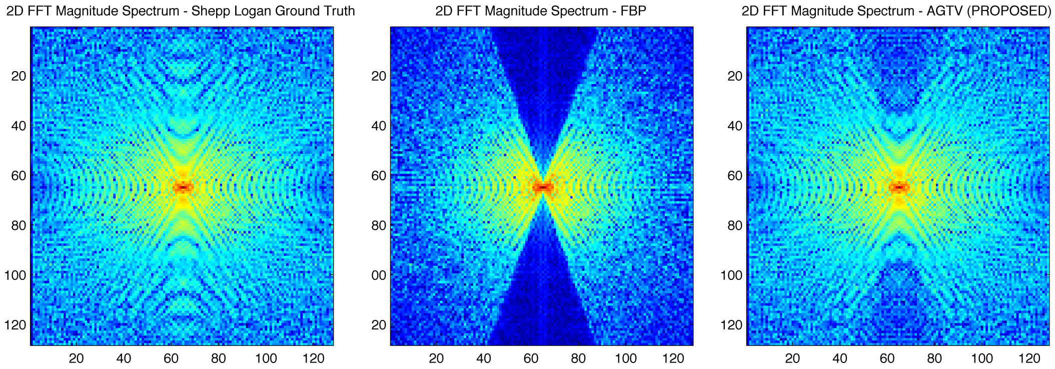


Figure 3.10: 2D FFT Magnitude components of the Shepp-Logan phantom, FBP reconstruction clearly showing the missing wedge and the AGTV reconstruction showing the extent to which the missing wedge can be recovered. The recovery is attributed to the CS-type nature of the sparsity exploiting image reconstruction method presented here.

Experiment 3.5: Tuning Hyper-parameters $(\lambda, \gamma, \mathcal{K})$

Aim

To study the variability of the three parameters that need to be tuned for every reconstruction

Methods & Conclusions

The reconstruction model presented has two hyper-parameters, λ for tuning the sparsity of CS based reconstruction and γ to tune the amount of smoothing and denoising in the reconstruction. While, these are model hyper-parameters and need tuning, the graph parameter $\mathcal{K}$, *i.e*, the number of nearest neighbors is quite easy to set for our application. This is shown in Fig. 3.11 where we perform a small experiment corresponding to the reconstruction of a 32×32 Shepp-Logan phantom from 36 projections $b \in \mathbb{R}^{36}$ using the pre-tuned parameters $\lambda = 0.1, \gamma = 5$ for different values of $\mathcal{K}$ ranging from 5 to 50. The results clearly show that the reconstruction is quite robust to the choice of $\mathcal{K}$, with a small error variation. Thus, $\mathcal{K}$ is easy to set for our application. As the complexity of our proposed algorithm scales with the number of edges $|\mathcal{E}|$ in the graph $\mathcal{G}$ and $|\mathcal{E}| \approx \mathcal{K}n^2$, it is recommended to set $\mathcal{K}$ as small as possible. However, a very small $\mathcal{K}$ might lead to many disconnected components in the graph $\mathcal{G}$. On the other hand, a very large $\mathcal{K}$ might increase the time required for the algorithm to converge and reduce the computational advantage we have over the NLTV method. Therefore, we choose to set $\mathcal{K} = 15$ for our experiments.

In order to show the variation of reconstruction error with (λ, γ) grid, we perform another experiment for the reconstruction of the Shepp-Logan phantom of size 32×32 from 36 projections. For this experiment we keep $\mathcal{K} = 15$ and perform the reconstruction for every pair of parameter values in the tuple (λ, γ), where $\lambda \in (0.1, 1)$ and $\gamma \in (0.1, 10)$. The reconstruction error grid is shown in Fig. 3.12. The minimum error 0.11 occurs at $\lambda = 0.2, \gamma = 0.1$. It is also interesting to note that the error increases gradually with an increase in the parameter values. Similar results were observed for larger image sizes. However, the best value of the parameters in heavily dependent on the image being reconstructed and the amount of noise in the projections. There is no hard and fast rule for tuning these parameters.

Experiment 3.6: AGTV with Real-Data

Aim

To investigate the effectiveness of AGTV with real Cryo-ET data from a biological specimen.

Method

A Cryo-ET data image was emulated in accordance with Experiment 2.5. A 256×256 cryo-ET image from open-sourse EMD database was used and 141 equally spaced and aligned 1D projections were used. The data was reconstructed using AGTV, in a simulation that ran for 35572 sec ($\approx$ 9.88 hours).

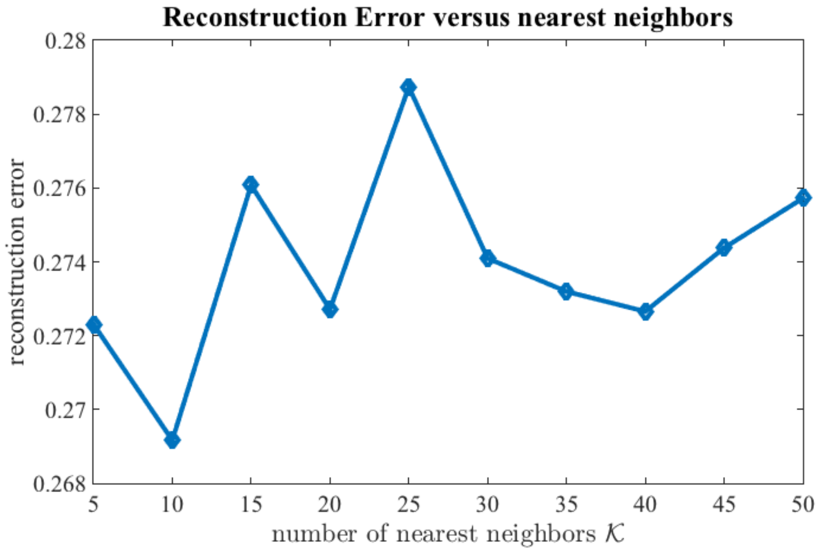


Figure 3.11: A small experiment corresponding to the reconstruction of a 32 × 32 Shepp-Logan phantom from 36 projections $b \in \mathbb{R}^{36}$ using the pre-tuned parameters $\lambda = 0.1, \gamma = 5$ for different values of $\mathcal{K}$ ranging from 5 to 50. The results clearly show that the reconstruction is quite robust to the choice of $\mathcal{K}$, with a small error variation.

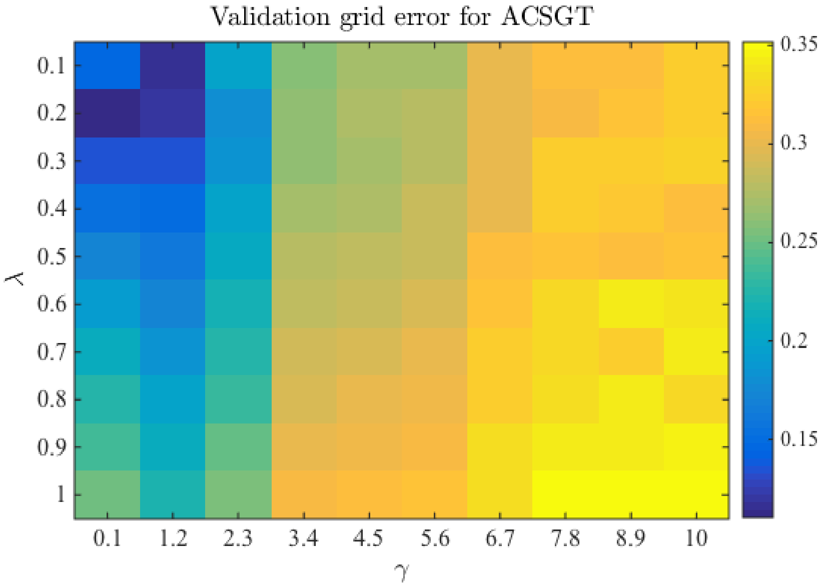


Figure 3.12: A small experiment corresponding to the reconstruction of a 32 × 32 Shepp-Logan phantom from 36 projections $b \in \mathbb{R}^{36}$ using the full parameter gird $\lambda = (0.1, 1), \gamma = (0.1, 10)$ for a fixed value of $\mathcal{K} = 15$. The minimum clustering error occurs at $\lambda = 0.2, \gamma = 0.1$. The results clearly show that the reconstruction error increases smoothly with the parameters.

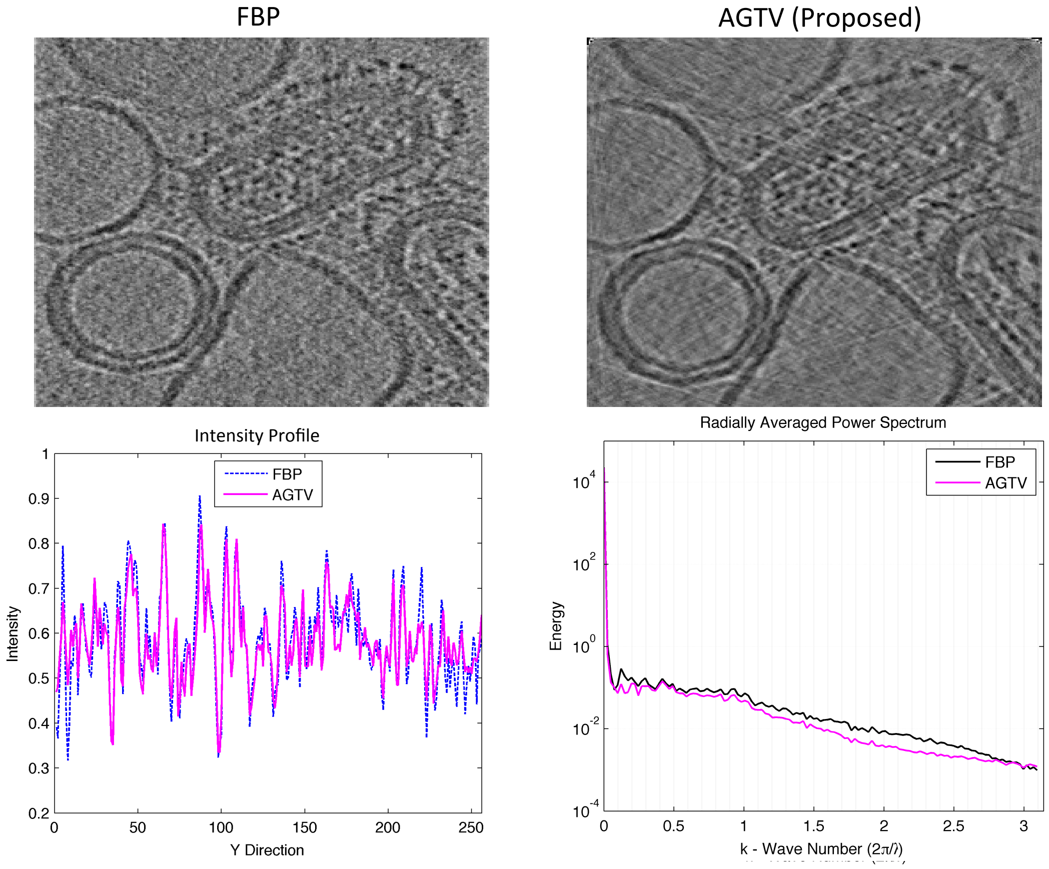


Figure 3.13: Figure showing results from FBP, AGTV reconstruction of Cryo-ET emulated data as well as their RAPS and intensity profiles. Data used is from an Influenza virus available from opensourse EM-Data-Base Entry: 4067. The image was reconstructed from 141 1D projections. As compared to sinogram denoising AGTV preserves more high frequency details.

In this case the ground truth (real result) is not known and the hyper-parameters can not be determined as accurately as for simulation experiments.

Conclusion

Fig. 3.13 shows the reconstruction, RAPS and the intensity profiles. It can be seen that when compared to the graph-sinogram denoising method presented in Fig. 2.11. this method behaves reasonably better. The intensity profiles and RAPS clearly shows that significantly more high frequency details are preserved. A further detailed analysis of the method with real data is beyond the scope of this thesis. More extensive parameter tuning may be required to achieve better results.

3.8 Shortcomings & Limitations

Complexity and Run-Time

The proposed AGTV method has proven to produce much better reconstructions as compared to the state-of-the-art CSTV method. Although, the proposed method is computationally far less cumbersome than the adaptive version of NLTV, it still suffers from a few problems which we discuss in this section. The computational complexity of the proposed method is $\mathcal{O}(I(J\mathcal{K}n^2+n^2\log(n^2)))$. As already presented in Algorithm 2, the method requires a double loop, the outer with I iterations and the inner with J iterations. For our experiments we usually set to $I = J = 25$ or $I = J = 50$. The main computational burden is offered by the graph construction, which needs to be performed every J iterations. Thus, the method still suffers from a high complexity because of the double loop and regular graph updates. The complexity of graph construction can be reduced by using a parallel implementation of FLANN which is provided by the authors [91]. The degree of parallelism can be increased at the cost of increasing approximation in the estimation of nearest neighbors. As a result of this the graph $\mathcal{G}$ will be different every time the FLANN algorithm is run. However, this does not effect the quality of the graph and for tomographic applications, negligible loss in the performance was observed. It is obviously of interest to reduce the number of inner iterations J and the complexity of the operations in the for loop. The results for the run-time have been given in Fig. 3.14. It can be seen that 256×256 reconstruction is already prohibitively complex on a rather modern general purpose processor. Any future work should therefore focus on introducing some approximations in the proposed algorithm to make it faster. The current setup acts as a proof-of-concept to motivate future studies along the lines of adaptive SEIR-based image reconstructions.

Hyper-parameter Tuning

Tuning the hyper-parameters is another short-coming of the proposed method. It is reasonable to set the number of $\mathcal{K}$-nearest neighbors to 10 or 15, however, the sparsity parameter λ and the GTV parameter γ need to be tuned properly and are not known beforehand. The results of the validation experiment from Fig.3.12 show that the error increases gradually with the parameter values. Any future study should also focus on finding smart methods to set these parameters automatically for specific tomographic applications. Parameter selection is a non-trivial problem and is an active research area among the theoretical computer science and mathematics community. Finding general methods and rules to tune such parameters is certainly not that simple and is usually accomplished by trial and error. This can be relatively less complex for tomographic modalities which image the same or similar object all the time, such as CT, but remains an integral problem for methods where the subject is significantly different every-time.

The Curse of Adaptivity

With regularization methods updated from data during iterations a major concern is whether incorrect features can be identified as reliable at some point and then reinforced thereafter. The possibility of this happening is higher when there is less data such as

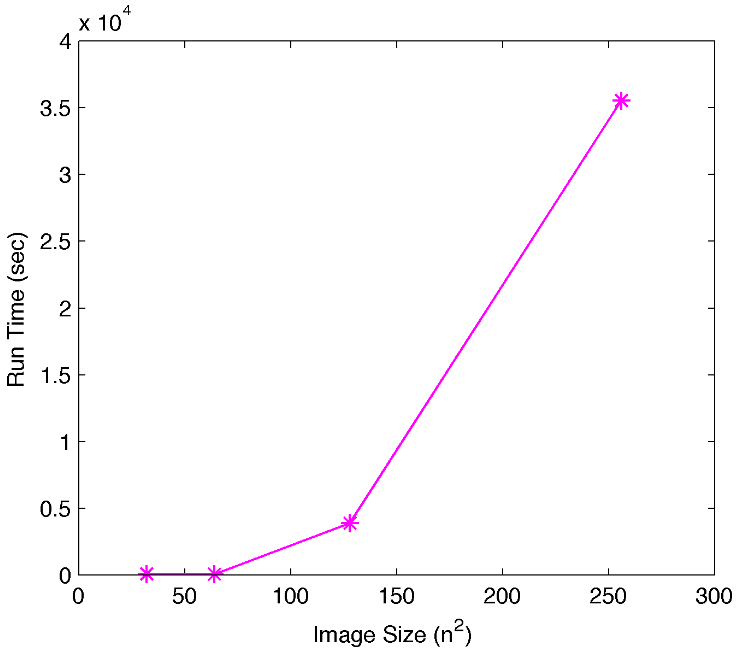


Figure 3.14: Figure showing run-time for AGTV corresponding to 2D tomography problems reconstructed from 1D data on an Intel Xeion Machine (E5450 12M Cache, 3.00 GHz, 128GB RAM).

when reconstructing from only 9 projections. This is due to the fact that less data will generate a highly erroneous FBP with many streaking artifacts which can be propagated to next stages of processing. This can be rectified by using a stronger filter on the FBP prior so the streaking or other artifacts can be reduced before a graph is constructed. Since the graph will be updated during every iteration the prior will continue to refine iteratively.

3.9 Conclusions

Similar to NLTV the proposed method (AGTV) goes beyond spatial similarity between different regions of an image being reconstructed by establishing a connection between similar regions in the image regardless of spatial distance. However, the presented approach is much more scalable and computationally efficient (when compared to adaptive NLTV-based SEIR methods) because it uses the approximate nearest neighbor search algorithm for graph construction, making it much more likely to be adapted for a real-data setting. However, as it stands now this approach is merely a proof-of-concept to motivate future studies which can make this method more efficient and tune parameters in a more smart way. The major unique point of the proposed approach is that it is adaptive. The non-local graph prior is updated every iteration making the connection between similar regions stronger. Thus improving the overall reconstruction quality as demonstrated by experiments. Since TV is a special case of graph TV the proposed method can be seen as a generalization of CS and TV methods and can promote future application specific studies for using CS-type SEIR methods with graph-based regularization for tomographic reconstruction from limited, noisy and erroneous data. Although this study was inspired by the Cryo-ET problem the method is general and may be adapted to a number of tomographic modalities. The proceeding chapter will focus on an approach with is much more computationally efficient when compared to AGTV and can provide robustness to noise in an intuitive way based on the geometric representation of data.

Chapter 4

Extended Field-based Noise Removal from Tomographic Reconstructions

4.1 Chapter Outline

This study[1] presents and analyses in detail an extended field-based tomographic reconstruction improvement paradigm. Unlike methods which locally or non-locally manipulate individual pixels of an image, extended field is based on the concept that extending the reconstruction space, which increases the dimensionality of the linear system being solved during reconstruction, facilitates the separation of signal and noise. A considerable amount of the noise associated with collected projection data arises independently from the geometric constraint of image formation, whereas the solution to the reconstruction problem must satisfy such geometric constraints. Increasing the dimensionality thereby allows for a redistribution of such noise within the extended reconstruction space, while the geometrically constrained approximate solution stays in an effectively lower dimensional subspace. Employing various tomographic reconstruction methods with regularization capability (ART, SIRT, Tikhonov, and Maximum Entropy Regularization) we performed extensive simulation and testing and we observed that enhanced dimensionality significantly improves the accuracy of the reconstruction. Through extensive empirical simulations we explicitly show that extending the reconstruction space reduces the error at a relatively lower regularization parameter thus allowing a better fit with the projections and preventing over-smoothing. Although the proposed method is used in the context of Cryo-ET, the method is general and can be extended to a variety of other tomographic modalities. This work has been published in [103].

4.1.1 Introduction

As summarized in Chapter one several reconstruction methods have been employed over the years, but generally they can be divided into two categories: Fourier slice theorem-based analytical methods and regularizing methods that are often iterative.

[1] Parts of this work specifically pertaining to results with real-data (Fig. 4.16) were completed as part of my lab rotation in the Structural Cellular Biology Unit at OIST between September-December 2012 before actually starting as an official PhD student in the lab. Parts of this work was also filed as a patent in 2013 [103].

Analytical algorithms, such as filtered back projection (FBP) [59] have been extensively used for tomographic reconstructions. However, such algorithms usually need a large number of projections, and it is not possible to incorporate *a priori* information and additional constraints into the approximate solution. When dealing with a reconstruction problem marred by missing and noisy data, regularizing methods are better suited. Among such methods, the mildly regularizing ART [60], SIRT [61], and its variants [62, 104] are often classified as Algebraic Reconstruction Methods (ARMs) or Row Action Methods (RAMs) [7]. Extended Field Iterative Reconstruction Technique (EFIRT) [105], which is based on ART and was published in 1974, briefly showed that reconstructing 2D images from 1D projections within an extended field can lead to low noise in the region of interest (ROI) and fast convergence. However, [105] offers only limited experimental and empirical evidence of the method. Algebraic methods have a limited regularization capability. This problem can be addressed by the use of variational regularization methods such as Tikhonov regularization [106] and Constrained Maximum Entropy Tomography (COMET) [15] that are equipped with explicit regularization and goodness-of-fit constraints.

COMET is an iterative reconstruction algorithm that performs a deconvolution of the point spread function (PSF) [12], and enhances the contrast and resolution specifically for Cryo-ET by increasing the SNR. This was first published in 1996 by *Skoglund et.al.* [15, 74], and has extensively been used for structure determination, e.g. in [107–109]. COMET can improve the fidelity of 3D reconstructions by reducing the shot noise. It increases (theoretically maximizes) the entropy relative to a prior obtained in the first iteration step from FBP, while iteratively deconvolving the PSF. In each iteration step it projects the new density to obtain virtual projections which are then compared to the original tilts using a Chi-squared goodness-of-fit statistic in order to find an optimal balance between relative entropy and goodness-of-fit.

Contributions

In 1974 *Crowther et.al.* [105] observed that extending the reconstruction space during ART reconstructions allows noise to spread out of the ROI. In this chapter we further build on this idea by contributing the following:

- With extensive empirical experiments we demonstrate that an extended field can go beyond ART and can be used to enhance a variety of reconstruction methods. We reconstructed several 2D phantoms from 1D projections corrupted with noise using ART (Kaczmarz, Symmetric Kaczmarz, Randomized Kaczmarz), SIRT (Cimmino, Landweber, DROP), and Tikhonov regularized reconstruction, and we observed that enhanced dimensionality resulting from a larger reconstruction space achieved higher correlation with the original phantom in every case.
- We verify how this method works by quantifying the amount of noise removed corresponding to the amount of noise added during simulations.
- We show that an extended field leads to better results when compared to non-extended reconstructions at a lower regularization parameter, thus preserving a better fit with the actual data and preventing over-smoothing.

- With extensive empirical simulations we demonstrate how extended field behaves at increasing extension steps.
- We further tested these effects on real Cryo-ET data, reconstructing the structure of colloidal silica using COMET and observed that an extended field renders low-noise reconstructions.
- We also explain the limitations of this method.

4.2 Preliminaries & Simulation Experiments

In order to analyze the *ill-posed inverse problem* of reconstructing data from projections, it is fundamental to have an accurate formulation of the *forward operator*, solving the problem of modeling the process of image formation in the absence of noise and measurement errors [11]. Mathematically, a discretized and simplified version of the noise-free forward operator can be described in terms of a linear system where projections b are collected from object x, given a matrix representation of the imaging device A (Recall Chapter 1).

$$Ax = b \qquad A \in \mathbb{R}^{m\times n} \;\; x \in \mathbb{R}^n \;\; b \in \mathbb{R}^m \tag{4.1}$$

The sinogram (S) represents the tilt series of raw data, *i.e.*, a matrix where each column represents a projection at a different angle. The vector b is the vectorized form of S. Each row of A corresponds to a single ray passing through the density being imaged. Since each ray only passes through a certain number of voxels, matrix A is usually sparse. The imaging model presented above represents a set of linear equations such that there is one equation for each ray passing through the object. Each equation here can be considered as a hyperplane in vector space, which can be defined as $\mathcal{H}_i = \{x \mid a_i^T x = b_i\}$. In an ideal, noise-free case, these linear equations would be consistent, and a solution would exist at the intersection of these hyperplanes. For the purpose of demonstration this has been shown for a system of two linear equations (Fig. 4.1a). However, in practice this is never the case, since the right hand side of the linear system is usually marred by noise $b = b^* + e$, where b^* is ideal data and e represents perturbation or data error. This renders the system inconsistent, and for algebraic techniques an approximate solution is sought within the region enclosed by the hyperplanes rather than at their intersection. Ill-posed inverse problems are generally not stable, *i.e.*, small perturbations in data can lead to large errors in the solution [18]. Hence, regularization methods are required to compute approximate solutions that are much less sensitive to perturbations in b [18]. Tikhonov regularization is a classical method of reducing the sensitivity of the solution to perturbations and noise in acquired data. The Tikhonov solution can be defined as the solution to the following problem, see [106, 110]:

$$\min_x\{\|Ax - b\|_2^2 + \lambda^2 \|x\|_2^2\} \tag{4.2}$$

$\|Ax - b\|_2^2$ measures the goodness-of-fit which essentially measures the effectiveness of the solution while the second term, $\|x\|_2^2$, measures the regularity of the solution,

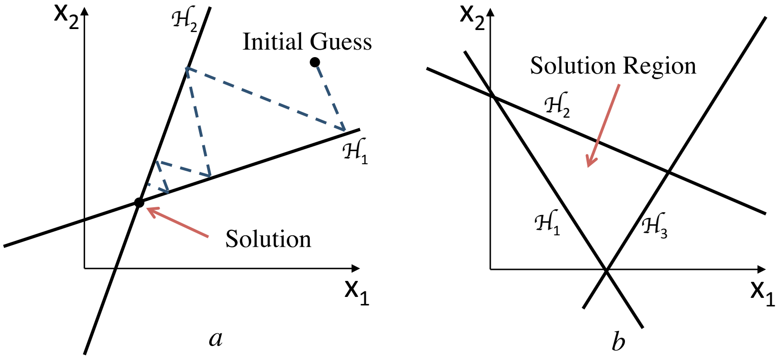


Figure 4.1: a) Consistent System of Equations having an exact solution, showing how Kaczmarz method approaches the solution starting from an initial guess. b) In case of an inconsistent noisy system the solution would be in a region bound by hyperplanes rather than at their intersection.

i.e., it controls the norm of x in order to suppress large noise components. The optimal balance between the two terms is weighted by the regularization parameter, λ. For a large λ, more weight is given to the minimization of the solution norm $\|x\|_2^2$ which then produces a more regular solution. For small values of λ the solution tends to be less smooth since more weight is given to fitting the noisy data.

4.2.1 Extended Reconstructions

Extended reconstruction essentially means reconstructing within a region larger than the region of interest (ROI). *This region outside the ROI gives extra freedom and flexibility to minimize data discrepancies and to maximize the consistency of the reconstruction within the ROI, with respect to the projections (Fig. 4.2).* Moreover, since we do not impose a non-negativity constraint in the extra region around the ROI, this essentially allows an extra degree of freedom when fitting the projections from our reconstruction to the sinogram. This concept can be better understood by thinking of noise as being limited by the projections, *i.e.*, the amount of noise associated with a specific reconstruction problem is constant. While the signal has a redundant occurrence, a clear representation, in every projection the noise is much less deterministic. If the reconstruction region is 'artificially' increased so as to provide an extra region for noise to distribute into, since it does not have a bound representation in all projections, there is a possibility that the noise in the reconstruction problem will be distributed over the entire reconstruction space thus spreading over the ROI and artificial region hence reducing the amount of noise in the ROI. This concept will be further explained in the proceeding sections and can be understood from Fig. 4.2.

4.2.2 ART with Extended Field (EART)

According to ART (Kaczmarz method [111]) the solution to a linear system of equations can be estimated starting from an initial guess and orthogonally projecting it onto successive hyperplanes until a stopping criterion is met (Figure 4.1a). However, in

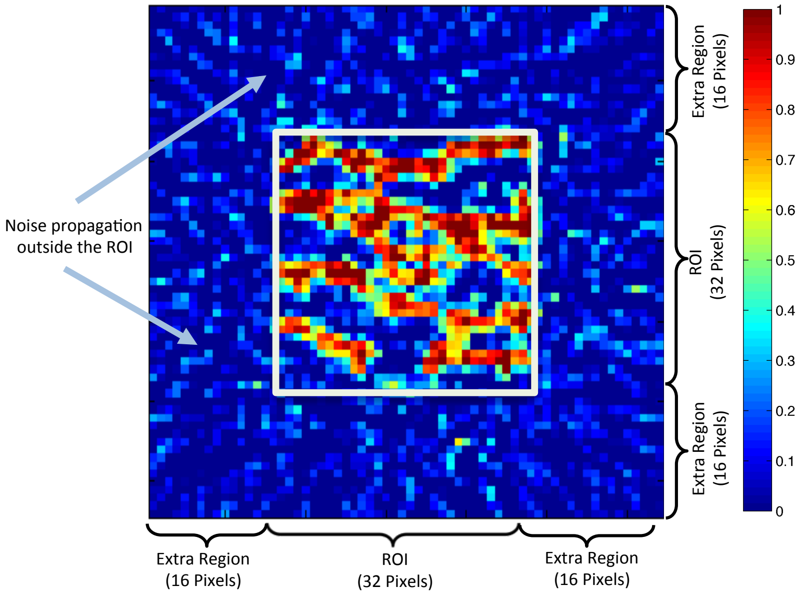


Figure 4.2: Reconstruction of a binary phantom with an extra region larger than the ROI. Unconstrained noise is redistributed out of the geometrically constrained ROI. The extra region gives flexibility to minimize data errors and maximize consistency of the solution within the ROI.

cases where the solution is inconsistent the introduction of a relaxation parameter γ can speed up convergence.

$$x^{k+1} = x^k + \gamma \frac{b_i - a_i^T x^k}{\|a_i\|_2^2} a_i \qquad \gamma \in (0, 2) \tag{4.3}$$

There are various formulations of Kaczmarz method such as symmetric Kaczmarz [41] and randomized Kaczmarz [31]. Most of these formulations are based on the way the rows, i, of matrix A are accessed. Extended ART performs a reconstruction within a larger reconstruction space, which is achieved by extending the sinogram S, either by choosing a large reconstruction region or via zero-padding. Extending the sinogram via zero-padding means padding each individual projection with zeros on each side (Fig. 4.4, Step 3). If a sinogram corresponding to a $N_1 \times M_1$ phantom is to be extended to a reconstruction corresponding to $N_2 \times M_2$, the sinogram is extended by padding each projection with $(\sqrt{M_2^2 + N_2^2} - \sqrt{M_1^2 + N_1^2})/2$ zeros on each side. Extending the sinogram from a reconstruction size of $N \times M$ to $(N + e) \times (M + e)$ corresponds to an increased dimension of the sinogram from $\sqrt{N^2 + M^2} \times n_\vartheta$ to $\sqrt{(N + e)^2 + (M + e)^2} \times n_\vartheta$ where n_ϑ is the number of projections.

Extending the sinogram increases the dimensionality of the linear system $Ax = b$ and makes vector b more sparse. Having an extra region outside the ROI allows inconsistencies and noise to spread out into the extra region, minimizing the discrepancy and enabling the solution in the ROI to be more consistent. This means that when reconstructing with ART, equation 4.3 is applied to not only the required ROI, but to an extra region outside, thus rendering a large reconstruction where the region of interest

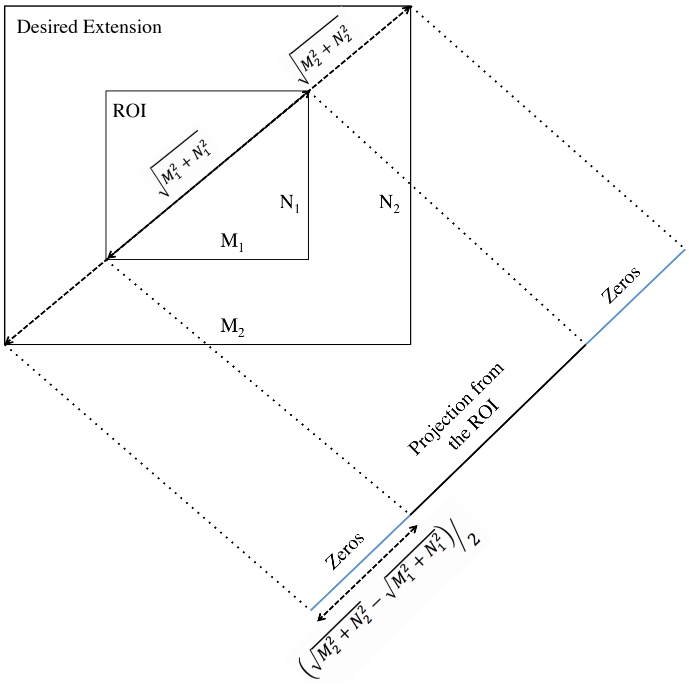


Figure 4.3: Image showing how the sinogram is extended to achieve an extra region outside the ROI. If a sinogram corresponding to a $N_1 \times M_1$ phantom is to be extended to a reconstruction corresponding to $N_2 \times M_2$, the sinogram is extended by zero-padding each projection with $(\sqrt{M_2^2+N_2^2} - \sqrt{M_1^2+N_1^2})/2$ on each side. For example, if the sinogram of a 64×64 phantom is extended to a 128×128 phantom the extension on each side of the sinogram would be 45.

Algorithm 3 ART (Kaczmarz Method)

```
INPUT: $S$,$k$,$A$, OUTPUT: $x^k$
***Vectorize S:*** $b = S(:)$
$x^0$=initial vector
**for** $k = 1, 2, 3...$ **do**
    $x^{k^0} = x^k$
    **for** $i = 1, 2, ..., m$ **do**
        $x^{k+1} = x^k + \gamma \frac{b_i - a_i^T x^k}{\|a_i\|_2^2} a_i$        $\gamma \in (0, 2)$
    **end for**
    $x^{K+1} = x^{k^m}$
**end for**
```

Algorithm 4 Extended ART (Kaczmarz Method)

```
INPUT: $S$,$k$,$A$,$Z$ OUTPUT: $x^k$
***Extend S in both directions:*** $S_{extended} = S + Z$
***Vectorize S:*** $b = S_{extended}(:)$
$x^0$=initial vector
for $k = 1, 2, 3...$ do
    $x^{k^0} = x^k$
    for $i = 1, 2, ..., m$ do
        $x^{k+1} = x^k + \gamma \frac{b_i - a_i^T x^k}{\|a_i\|_2^2} a_i \qquad \gamma \in (0, 2)$
    end for
    $x^{K+1} = x^{k^m}$
end for
```

is preserved and noise has relatively more room in which to be distributed. On removing the extra region, inconsistent discrepancies are removed and the consistent solution is isolated. The density in the extra region originating from inconsistencies, noise, and fringe effects can be positive or negative; hence, the non-negativity constraint often used during ART should not be applied in the extra region.

4.2.3 Simulations with EART

Experiment I: To test the procedure defined in the previous subsection a 32 × 32 binary phantom [31] (x^*) was used. A sinogram was created by taking projections at every 5 degrees starting from 1° to 180° (*i.e.*, $\theta = 1 : 5 : 180$) and projections were corrupted with normalized random noise (relative level = 0.05). This sinogram was then reconstructed using all three variants of ART. A version with an extended field (EART) was also reconstructed using all three ART procedures. The EART reconstructions were performed by increasing the dimensions of the sinogram corresponding to a 64×64 reconstruction. The error rate for the k^{th} iteration relative to the original phantom was calculated using $\left\|x^k - x^*\right\|_2$, where x^* is the original phantom and x^k is the result of the k^{th} iteration of the iterative method being used. The SNR for the reconstructions can be estimated using $SNR = \|x^*\|_2 / \left\|x^k - x^*\right\|_2$. It should be noted here that the error rate or SNR for extended reconstructions is always calculated corresponding to the ROI excluding the extended region. The complete experimental setup is shown in Fig. 4.4.

Experiment II: A 64 × 64 Shepp-Logan phantom [26] used to further test the arguments presented at a higher noise level and the projections in the sinogram were corrupted by normalized noise of relative level 0.1. The EART reconstructions were performed by increasing the dimensions of the of the sinogram corresponding to a 128 × 128 reconstruction. The rest of the simulation setup is exactly the same as experiment I.

Traditionally, Kaczmarz method and its variants show a semi-convergence behavior, where the first few iterations show a sharp decrease in error rate and then slow down, or in certain cases, diverge from the optimal solution. From the simulations it is evident that EART follows a similar trend, but tends to achieve a lower error rate

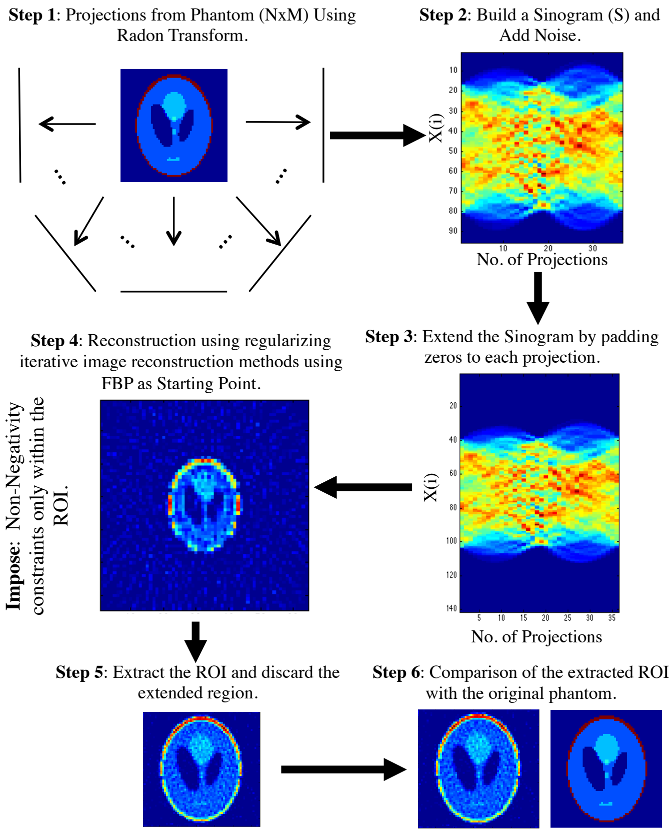

Step 1: Projections from Phantom (NxM) Using Radon Transform.
Step 2: Build a Sinogram (S) and Add Noise.
X(i)
10
20
30
40
50
60
70
80
90
10
20
30
No. of Projections
Step 3: Extend the Sinogram by padding zeros to each projection.
X(i)
20
40
60
80
100
120
140
5
10
15
20
25
30
35
No. of Projections
Step 4: Reconstruction using regularizing iterative image reconstruction methods using FBP as Starting Point.
Impose: Non-Negativity constraints only within the ROI.
Step 5: Extract the ROI and discard the extended region.
Step 6: Comparison of the extracted ROI with the original phantom.

Figure 4.4: Flow diagram showing the simulation setup for various extended field simulations with several different iterative reconstruction methods. It should be noted here that in step 3 the sinogram is zero-padded in the direction of the projections *i.e.* adding an extra region to each projection.

relatively faster than ART (Figure 4.5). Although EART needs more iterations to realize its full potential, it gives a lower error at the semi-convergence point of ART. While ART starts diverging, EART continues to reach semi-convergence in a later cycle, with a lower error-rate. These two experiments at relatively low and high noise, respectively, both show similar behavior and indicate that increasing the dimensionality of the reconstruction problem by increasing the reconstruction space has a direct effect towards improving reconstruction quality. Fig. 4.6 shows an intensity profile of a single line through each of the phantoms in experiments I and II. The difference between FBP, ART, and EART is clearly evident. Simulation results with a much higher error rate (relative level $= 0.20$) have been summarized in Table I.

Experiment III: In order to check the effect of the relaxation parameter γ, we performed several ART and EART reconstructions with varying γ between $0 < \gamma < 2$ (Fig. 4.7). For both phantoms, the error rate is lower for EART reconstructions than for ART reconstructions, except at very low γ values, when the relative noise is relatively high.

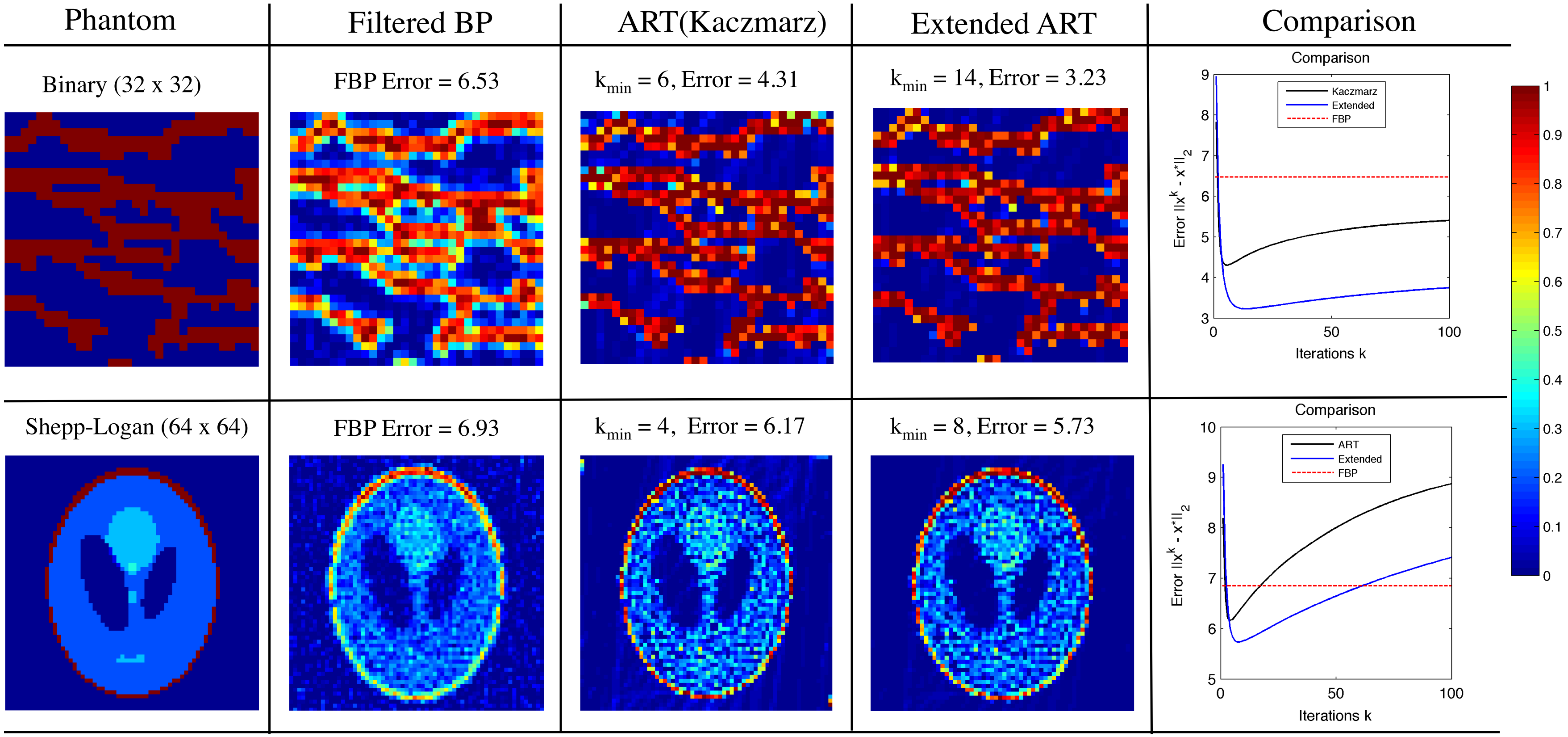


Figure 4.5: Comparative analysis of reconstructing a binary (relative noise = 0.05) and Shepp-Logan (relative noise = 0.1) phantom with ART and EART shows that EART can achieve better results with a faster semi-convergence rate and lower error rate than ART. Codes to generate these figures are available with this thesis.

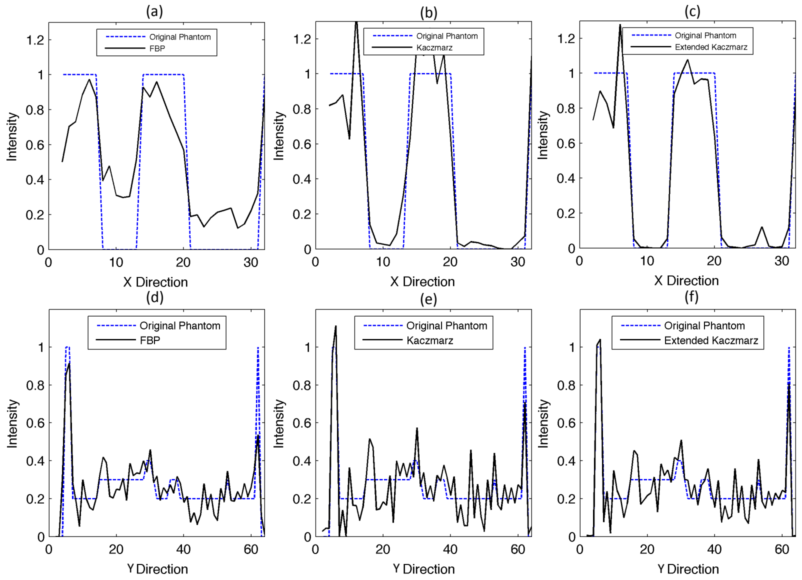


Figure 4.6: a-f)Intensity profiles of a line through the original and reconstructed binary and Shepp-Logan phantoms using FBP, ART (Kaczmarz) and EART (Extended Kaczmarz). The EART reconstructions show a relatively better fit to the original phantom as compared to EART reconstructions.

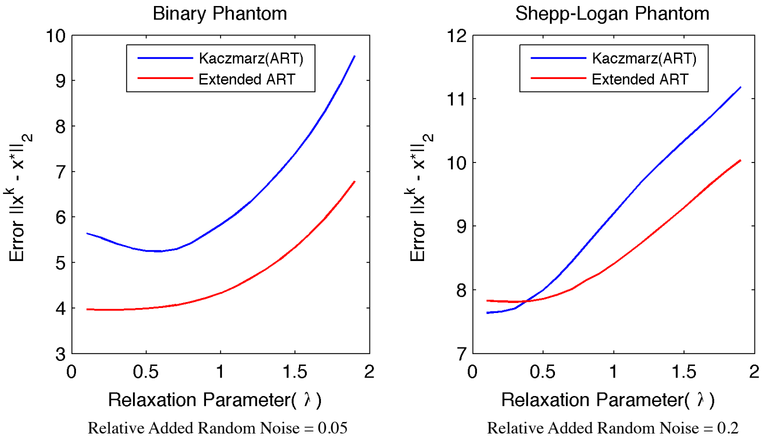


Figure 4.7: Comparative analysis of ART and EART with respect to changing relaxation parameter γ.

Table 4.1: Comparison of error rates for ART and EART using various phantoms at 20% noise

Phantom	FBP	Kaczmarz	Ex. Kaczmarz	Sym. Kaczmarz	Ex.Sym. Kaczmarz	Rand. Kaczmarz	Ex.Rand. Kaczmarz
Smooth	29.13	27.54	25.91	27.03	24.97	25.36	24.23
Shepp-Logan	13.98	10.21	9.78	10.32	9.61	10.16	9.01
Binary	11.04	8.40	7.51	8.98	7.83	8.36	7.12
Four-Phase	19.24	16.31	15.41	19.16	15.39	19.49	14.98

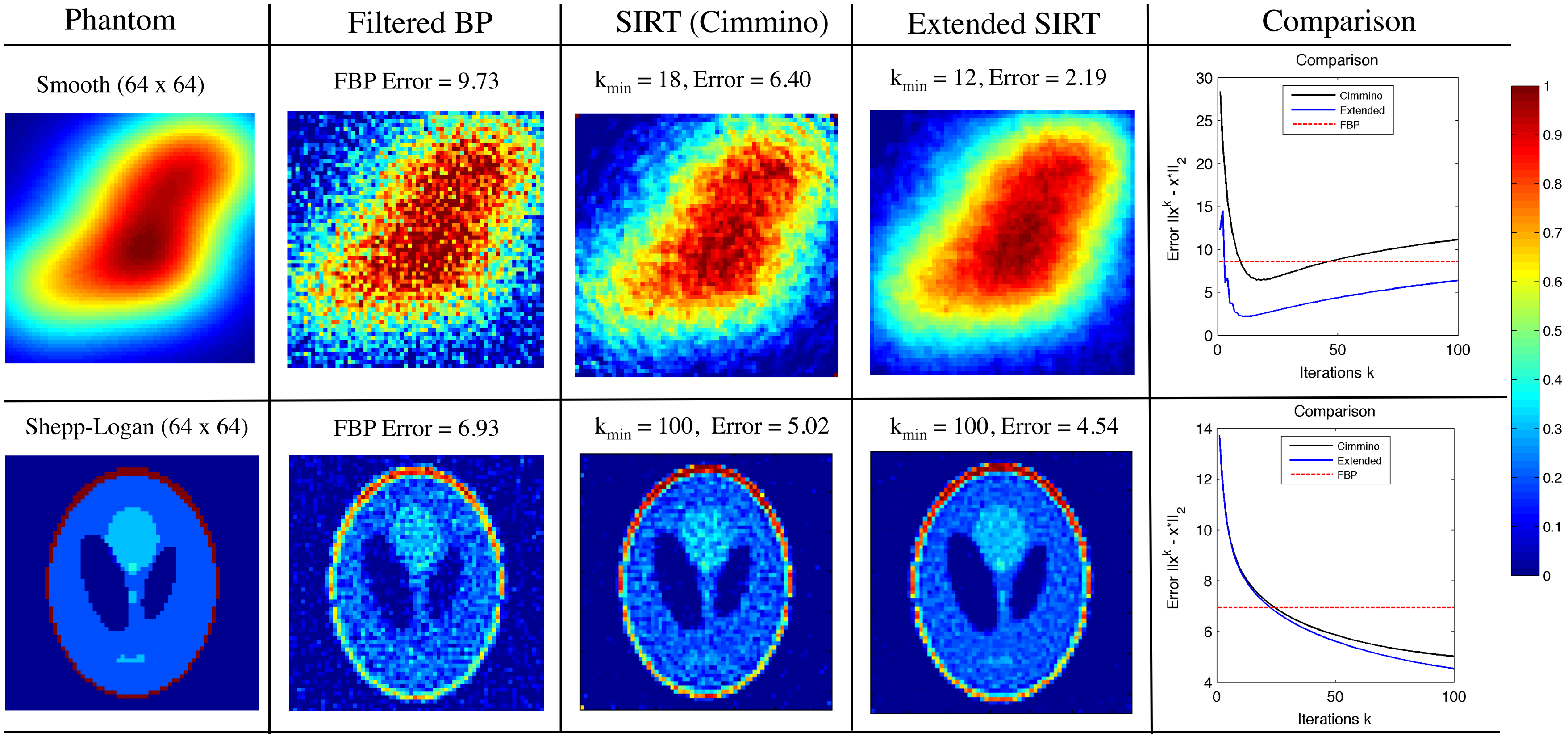


Figure 4.8: Comparative analysis of reconstructing a smooth phantom (relative noise = 0.05) and Shepp-Logan (relative noise = 0.1) with SIRT and ESIRT shows that ESIRT can achieve better results with a faster convergence and lower error rate as compared to SIRT.

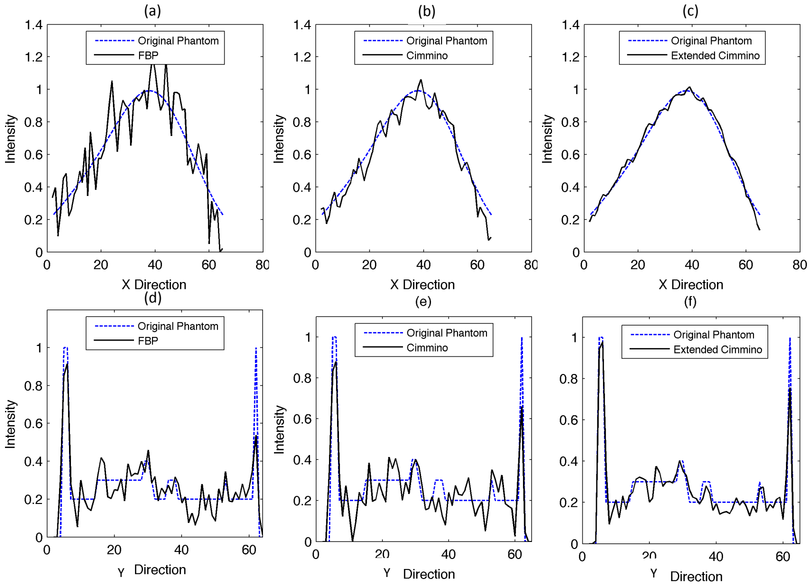


Figure 4.9: Intensity profiles of a line through the original and reconstructed smooth and Shepp-Logan phantoms using FBP, SIRT (Cimmino) and ESIRT (Extended Cimmino). The ESIRT reconstructions show a better fit to the original phantom as compared to SIRT reconstructions.

Table 4.2: Comparison of error rates for SIRT and ESIRT using various phantoms at 20% noise

Phantom	FBP	Cimmino	Ex.Cimmino	Landweber	Ex.Landweber	DROP	Ex.DROP
Smooth	29.13	23.26	21.70	23.43	20.11	23.41	20.32
Shepp-Logan	13.98	9.13	8.34	9.13	8.33	8.78	7.82
Binary	11.04	7.18	6.32	7.38	6.34	7.09	6.28
Four-Phase	19.24	13.96	12.44	13.77	12.19	13.63	11.91

Table 4.3: Comparative analysis of error rates for Tikhonov and extended Tikhonov reconstructions at varying relative noise and number of projections.

Phantom	Relative Noise	5%			10%			20%			50%		
	No. of Proj.	**36**	**18**	**9**	**36**	**18**	**9**	**36**	**18**	**9**	**36**	**18**	**9**
Smooth	Regular	5.19	9.46	11.31	12.87	14.18	16.63	18.61	27.49	34.54	68.46	77.38	82.37
	Extended	2.06	3.13	5.62	4.21	4.56	5.13	9.34	10.95	12.87	19.87	23.88	26.19
Shepp-Logan	Regular	3.98	4.35	5.28	4.78	7.34	8.94	8.17	10.46	16.56	17.74	20.71	32.84
	Extended	3.18	3.87	4.84	4.21	4.50	5.35	5.21	5.25	5.16	6.86	7.07	9.04
Binary	Regular	3.21	4.29	7.18	4.67	6.11	8.24	7.32	12.45	14.91	19.20	24.12	29.11
	Extended	2.46	2.98	3.30	3.09	4.12	4.36	6.71	8.43	9.88	11.61	14.28	18.29
Four-Phase	Regular	8.99	10.97	12.46	10.38	16.38	19.03	12.28	22.43	27.21	38.94	44.03	49.08
	Extended	6.32	7.11	8.40	9.11	10.49	12.26	11.90	15.38	18.43	22.19	26.46	29.15

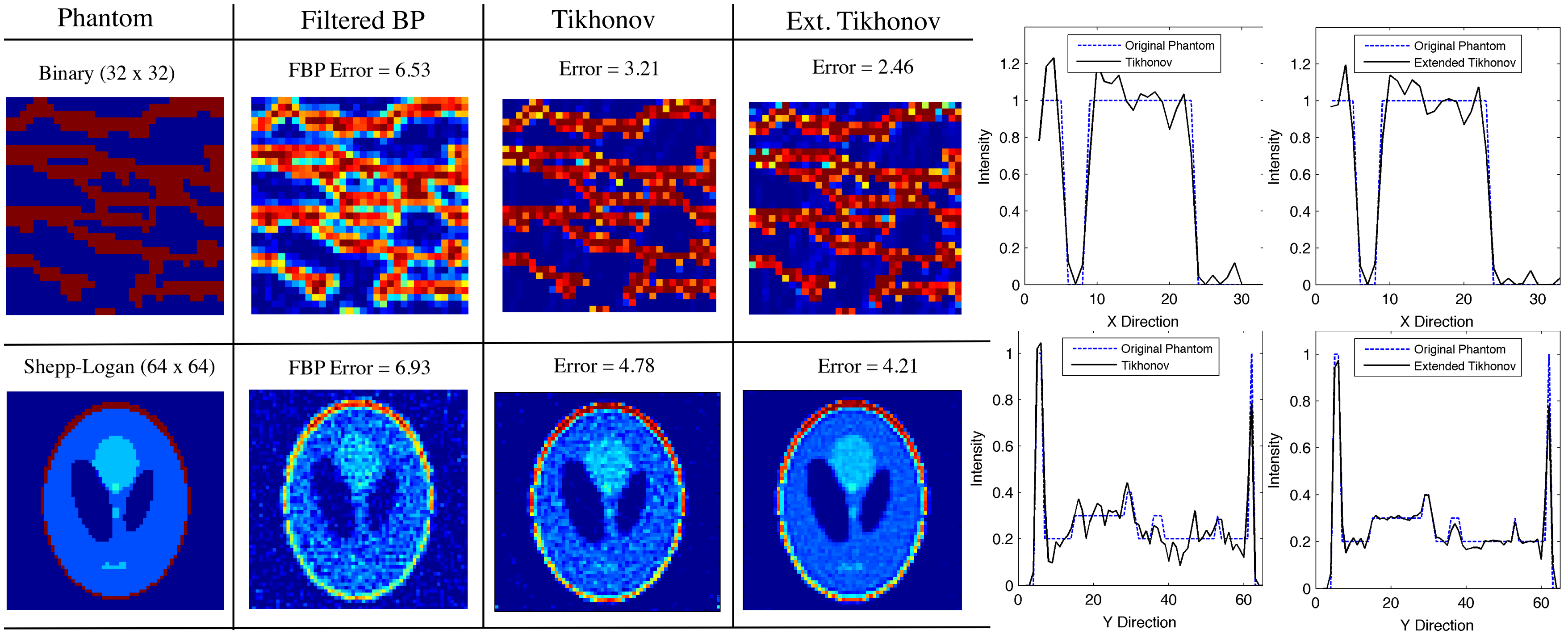


Figure 4.10: Comparative analysis of reconstructing a binary (relative noise = 0.05) and Shepp-Logan (relative noise = 0.1) phantom with Tikhonov and extended Tikhonov regularization. Extended Tikhonov performs better than regular Tikhonov regularization. Intensity profiles of for both phantoms show that extended Tikhonov performs better than Tikhonov and has a higher regularization effect.

4.2.4 SIRT with Extended Field (ESIRT)

Unlike ART the class of SIRT methods are "simultaneous" which means that all rows of matrix A are used simultaneously in every iteration [31]. SIRT can thus perform faster than ART. It also has better regularization properties as compared to ART. The SIRT class of methods can be generally defined as follows:

$$x^{k+1} = x^k + \gamma_k T A^T M(b - Ax^k) \tag{4.4}$$

Here x^k and x^{k+1} denote the current and successive iteration respectively. γ_k is the relaxation parameter, and the matrices M and T are symmetric positive definite. Various methods of SIRT have different M and T [112][113]. Ideally, the iterations mentioned in equation 9 converge to a solution x^*. However, similar to ART this is usually not the case and the solution converges to a region and oscillates in this region over successive iterations. The class of SIRT methods also follow a semi-convergence behavior as shown in [114, 115]. Similar to EART a version of SIRT can be formulated by extending the region outside the ROI either by zero-padding or within the region being reconstructed. This will lead to the enhanced dimensionality of the linear system being solved and extra reconstruction space, these extra dimensions act as a mechanism for segregating most of the background noise.

Simulation and testing was performed on three variants of SIRT, *i.e.*, Landweber, Cimmino and Diagonally Relaxed Orthogonal Projections (DROP). Landweber's method [116] simply originates from equation 9 where $M = I$ and $T = I$. Cimmino's method [62] works on the premise that each iteration is the weighted average of the projections of the previous iteration on all the hyperplanes:

$$x^{k+1} = x^k + \gamma_k \frac{1}{m} \sum_{i=1}^{m} w_i \frac{b_i - \langle a_i, x^k \rangle}{\|a_i\|_2^2} a_i. \tag{4.5}$$

In terms of equation 9 Cimmino is defined as having $M = D$ and $T = I$, where D can be defined as follows:

$$D = \frac{1}{m} diag\left(\frac{w_i}{\|a_i\|_2^2} a_i \right). \tag{4.6}$$

DROP is an extension of Cimmino's original method and incorporates information about the sparsity of matrix A [29]. When extending the dimensionality by adding an extra region outside the ROI the vector b is increased and becomes sparse. The extra region adds additional hyperplanes to the system passing through the origin. This essentially means that for Cimmino's method more hyperplanes are averaged to achieve the output of every iteration which in turn leads to distribution of noise to extra space. Similar to EART the solution stays in an effectively lower dimensional sub-space and the noise redistributes to higher dimensions.

4.2.5 Simulations with ESIRT

A 64×64 smooth [31] and Shepp-Logan [26] phantom was reconstructed using FBP, SIRT, and ESIRT with all three variants mentioned in the previous sub-section. Sinograms for these phantoms were created by taking projections every 5 degrees from 1° to 180° degrees (*i.e.*, $\theta = 1 : 5 : 180$) and corrupted with normalized noise relative level 0.05 for smooth and 0.1 for Shepp-Logan, respectively. ESIRT reconstructions were performed by increasing dimensions of the sinogram corresponding to a 128×128 reconstruction. Both iterative methods were stopped at 100 iterations (k). The error rate was calculated as mentioned in the previous subsection. Fig. 4.8 shows a comparative analysis of SIRT and ESIRT methods in detail. Visual inspection of the reconstructed phantom clearly shows that ESIRT has better regularization properties than SIRT. It is evident that ESIRT tends to converge faster than SIRT and achieves a much better correlation with the original phantom. In the case of ESIRT, the divergence from semi-convergence is also slower. Fig. 4.9 shows an intensity profile of a single line through SIRT and ESIRT phantoms reconstructed using Cimmino's method and clearly indicates that ESIRT behaves much better than SIRT. Simulations at a high noise level (relative level = 0.20) are presented in Table 4.2.

4.2.6 Simulations with Extended Tikhonov Regularization (ETR)

Tikhonov Regularization (TR) has already been explained in earlier. TR has a higher regularization capability than SIRT and ART and has been extensively used in the context of various *ill-posed inverse problems* [18]. We conducted several experiments to test the effectiveness of an extended field outside the ROI while reconstructing a density with TR.

We reconstructed a 64×64 Smooth phantom (relative noise = 0.05) and a 64×64 Shepp-Logan (relative noise = 0.1) using both TR and ETR. ETR simulations were conducted by zero-padding the sinogram corresponding to a 128×128 reconstruction (Fig. 4.10). In line with our findings for ART and SIRT, ETR performed better than TR when tested on a variety of phantoms under different noise conditions, as well as with varying numbers of total projections (Table 4.3). It is evident from these results that TR regularizes more powerfully than the methods discussed in the previous sections; hence, in agreement with the underlying heuristics, the improvement due to the use of extended field technique is also significantly larger.

Additional Experimental Observations

It should be noted here that decreasing the field instead of extending it leads to having a reconstruction space smaller than corresponding projection size. On simulating several reconstructions by decreasing the field we observed that it always leads to a more erroneous reconstructions. This is the case if the error rate is calculated relative to the original phantom without excluding the region which was reduced by having a smaller field of view. That said, an extended field does not require an object to be fully enclosed in the ROI. If the object is not enclosed, *i.e.*, the sinogram is modified to reconstruct a part of the density the projections still hold information about the removed part and the reconstruction / regularization method used will treat it as noise. On reconstructing

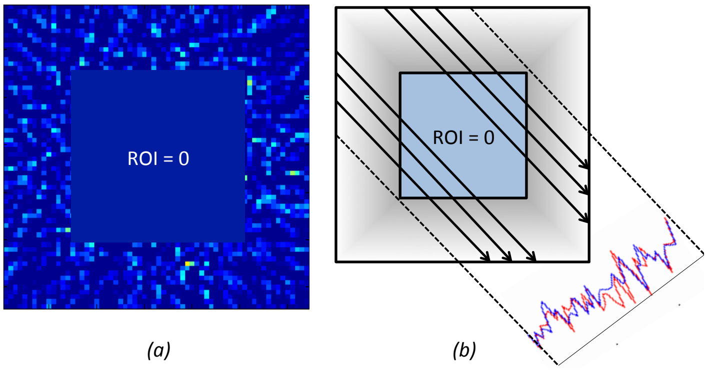


Figure 4.11: Experimental design to quantify the noise in the extended field. The pixels in the reconstructed ROI are set to zero and projections are taken relative to the ROI and compared with noise added to individual projections of the initial sinogram.

partial densities from four different phantoms at different noise levels we observed that an extended field always improved the error rate. This is the case if the error rate is calculated relative to the part of the phantom which was reconstructed. The reason why FBP has lower error than the starting point of algebraic methods is because the FBP output is filtered before the error rate is calculated while the unfiltered version is used as a prior for algebraic methods.

4.3 Quantifying the Amount of Noise Removed

In this section we empirically analyze and quantify the amount of noise that redistributes into the extended region. This is essential to verify the effectiveness of extended reconstructions and allows us to quantify the amount of noise removed compared to the amount of noise added.

Experimental Setup: 36 projections were generated from a 64×64 Shepp-Logan Phantom and the projections were corrupted with 15% added normalized random noise. The noisy sinogram was extended by zero-padding corresponding to a 128×128 reconstruction space. The resulting sinogram was reconstructed using ART (Kaczmarz), SIRT (Cimmino) and Tikhonov using FBP as the starting point. The 64×64 ROI in the resulting 128×128 extended reconstruction is set equal to zero (Fig. 4.11). We then take 36 projections equivalent to the size of the ROI from this matrix (Figure 4.11-b). This essentially means that we summed up pixel-by-pixel, all the noise removed from the ROI while leaving out regions that did not contribute to the original projections. These projections were then compared to the original noise added to each projection (Figure 4.12).

The good fit between the projected extended field region (blue curve in Figure 4.12) and the noise added to the projections (red curve in Figure 4.12) shows that most of the noise was redistributed to the extended region during regularization. Table 4.4 shows the average correlation between red and the blue signals. The fit for Tikhonov is better than SIRT and ART and the fit for SIRT is better than ART. This clearly indicates

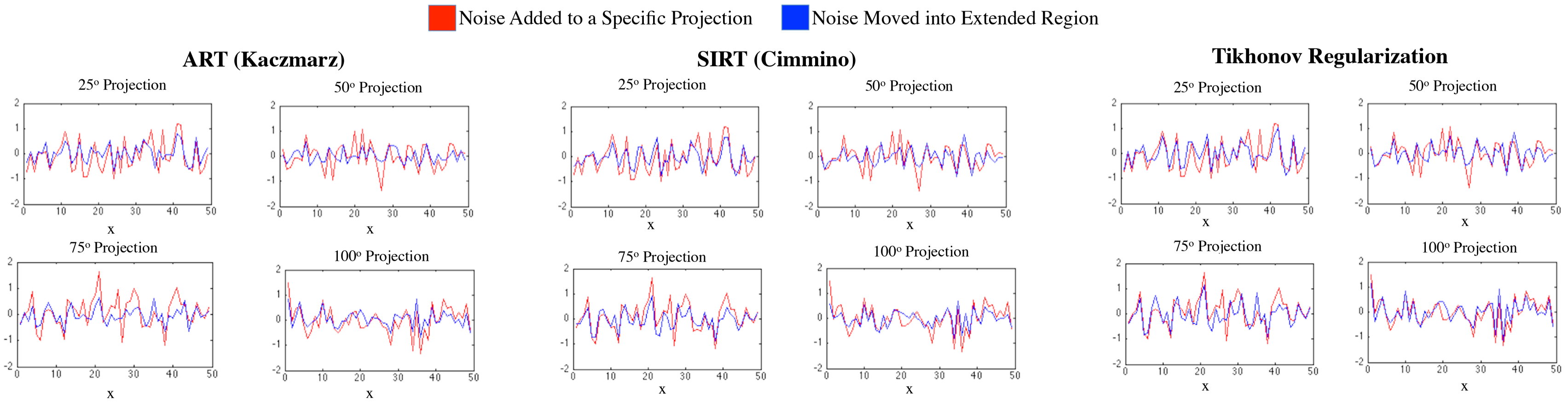


Figure 4.12: Comparison of noise removed, *i.e.*, noise redistributed into the extended field and the noise originally added to the projection at 25^o, 50^o, 75^o and 100^o for extended ART, SIRT and Tikhonov reconstructions of a Shepp-Logan phantom. The good fit shows that the amount of noise removed by the flexibility provided by the extended field is approximately equivalent to the amount of noise added while conducting the simulation (Table 4.4). Results for the entire sinogram and similar results for ART and SIRT have been presented in the supplementary material.

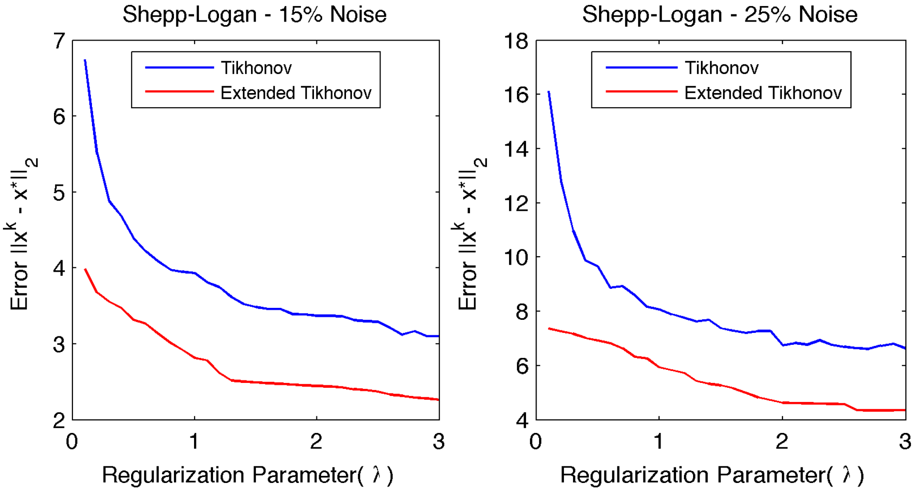


Figure 4.13: Comparative analysis of error rates for Tikhonov and Extended Tikhonov reconstructions at a varying regularization parameter ($0.1 \leq \lambda \leq 3$). The error rates are calculated corresponding to the ROI for Extended Tikhonov reconstructions. It can be clearly seen that Extended Tikhonov achieves a much better error rate at a lower λ than regular Tikhonov.

that Extended Field works more powerfully for methods with a higher regularization capability. Further results from similar simulations with SIRT, ART and Tikhonov are elaborated in the supplementary material.

4.4 Enhanced Regularization Using Extended Field

It is evident from experiments with algebraic methods (ART, SIRT and their variants) as well as with variational regularization (Tikhonov) that the extended field directly enhances the effect of regularization. It can also be clearly seen that extended field works better with stronger regularization methods and at higher noise levels e.g. extended field with Tikhonov regularization at 50% noise can improve the reconstruction several folds (Table 4.3, column 4). The distribution of the noise removed from Tikhonov reconstructions also fits much better with the actual noise added as compared to ART and SIRT (Table 4.4).

We reconstructed various phantoms using Tikhonov regularization for a changing regularization parameter (λ in equation 3) and observed that extended field reconstructions can achieve an enhanced regularization effect when compared to regular reconstructions at a relatively lower λ (Fig. 4.13). In accordance with our previous claims the rationale behind this is the unique ability of an extra region to enhance the consistency of the solution with respect to the sinogram.

The regularization parameter controls the balance between the fit and the regularity of the solution. A large λ produces a strongly regular solution i.e. the solution is smoother. A lower λ produces a solution more faithful to the data (projections).

Since having an extended field reduces the error at a relatively lower regularization parameter, this means that it can produce a solution which is more faithful to the data while preventing over-smoothing. The extended region allows for an increased fit between the projections (sinogram) and the reconstruction by achieving a better error rate at a lower λ.

4.5 Optimal Extension Size

Simulations in the previous sections show how an extended field behaves with a variety of reconstruction methods, at various noise levels and at different number of projections. We clearly show that extended field can have a promising effect on reconstructions by lowering the error rate and improving the correlation with the actual phantom. However, extending the sinogram to achieve an extra reconstruction space means increasing the size of the reconstruction problem which has a computational constraint. Keeping this in mind, it is necessary to estimate an optimal extension size especially when dealing with real data.

We reconstructed 64×64 Shepp-Logan and Binary phantoms at varying extension steps $(0, 2, 4, 8, ...512)$ and at different noise levels. Apart from these the experimental setup is the same as mentioned in the previous section. Fig. 4.14 summarizes these numerical tests and suggests that all extended reconstructions show a sharp decrease in the error rate during the first few extension steps followed by a saturation in error improvement. As the extended field is increased more and more noise can be removed from the ROI into the extra region. However, the extended field can not continue to improve the ROI beyond a certain point, this is due to the fact that it can not remedy the noise and inconsistence originating from missing data (e.g. fewer projections, missing wedge). It can be clearly seen from the experiments that at higher noise levels the sharp decline in the error rate during the first few extension steps becomes much more prominent. Also, error rates for stronger variational regularization methods (Tikhonov) show a sharper decline when compared to mildly regularizing algebraic methods (ART, SIRT). After saturation the error rate stabilizes and varies only slightly despite large extension sizes.

Extremely large extension steps lead to a slight increase in the error rate. For example extending a 64×64 Shepp-Logan phantom corresponding to a 4096×4096 reconstruction gives an error of 4.268 as compared to a 256×256 reconstruction which yields an error rate of 4.257. Moreover, the extended field reconstruction of phantoms with no noise added leads to extremely slight improvements in the error rate, as mentioned inconsistencies due to missing data can not be corrected, the slight improvement comes form the correction of errors originating from the numerical treatment of the problem.

The optimal extension size certainly depends on the type of raw data, the amount of noise and the regularization capability of the reconstruction method. As shown by our experiments a strong regularization method will definitely need a smaller extension to remove most of the noise from the ROI into the extended region. For instance, reconstructions in [105] were performed using ART, which has mild regularization ca-

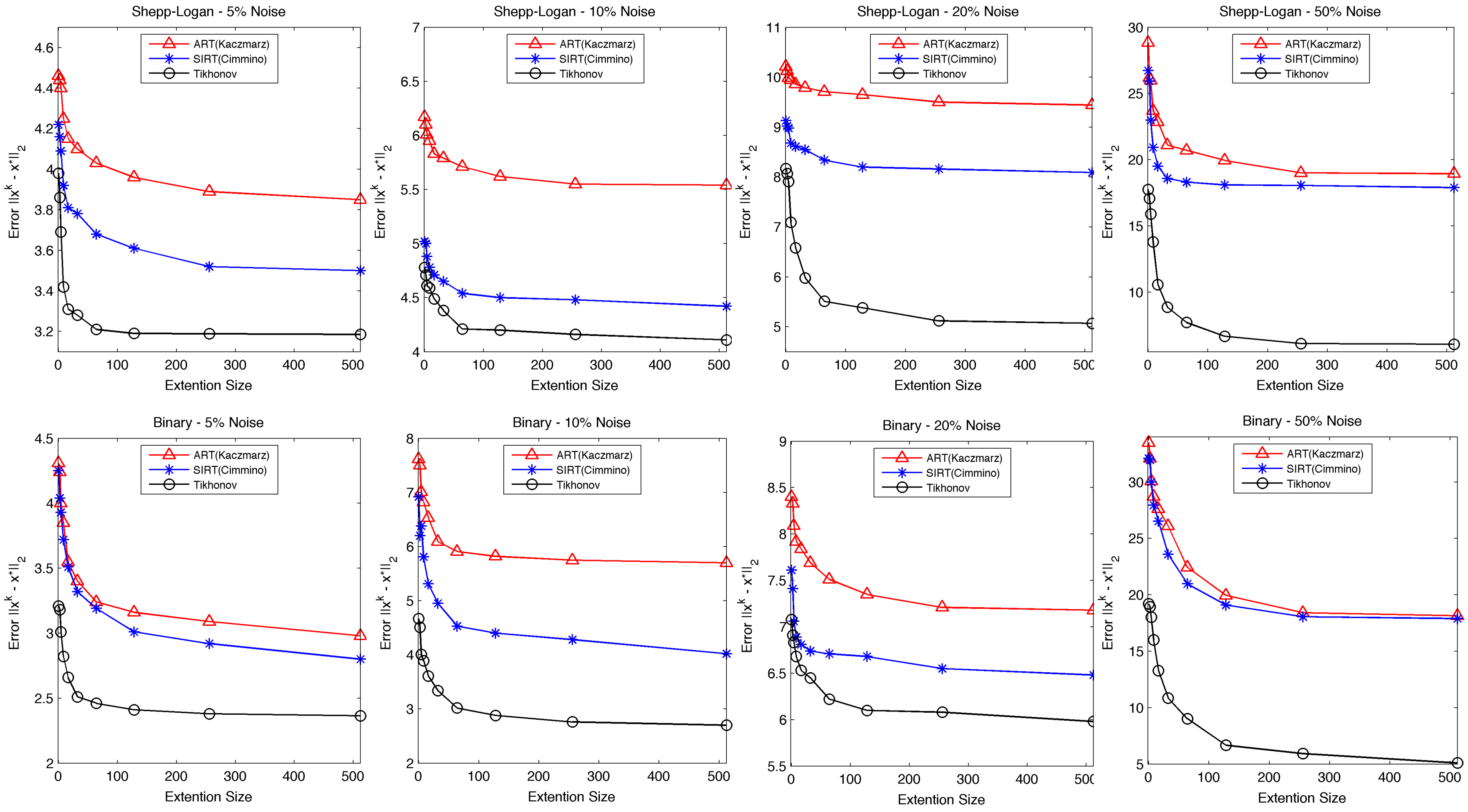


Figure 4.14: Comparative analysis of error rates for ART, SIRT and Tikhonov reconstructions at varying extension steps $0, 2, 4, 8, ...512$ starting from a 64×64 phantom. It can be clearly seen that most of the benefit of an extended field is achieved if the size of the extension is half the actual reconstruction size. All reconstructions show a sharp improvement during the first few small extension steps followed by a saturation of error improvement for larger extensions.

pabilities, and in that specific case twice the reconstruction size may have been needed. When using Tikhonov regularization an extension of half the size of the actual phantom, removes approximately 80% of the maximum possible noise redistribution. Contrary to [105] we observed that extending the reconstruction space twice or more the actual reconstruction space is not necessary to realize the full potential of an extended field. The ratio of error improvement to extension size decreases at higher extensions. It is computationally inefficient to increase the extension size beyond half the size of the actual reconstruction if powerful enough regularization method is employed, since the gain in SNR beyond that is not significant.

The number of projections also play a role in determining the most optimal extension size. This was verified by an additional experiment. We reconstructed a 64 × 64 Shepp-Logan phantom projections corrupted with 15% and 25% random normalized noise with different number of projections and varying extension sizes using Tikhonov regularization. The results are shown in Figure 4.13. We can clearly see that a reconstruction with fewer projections saturates faster when compared to a reconstruction with more projections. This makes sense because increased number of projection will need extra variables outside the ROI to satisfy the noise introduced by relatively more projections.

In short, the optimal and most computationally efficient extension size can be decided depending on the on the type of data being reconstructed, the amount of noise in the data, the number of projections and more importantly the regularization capabilities of the reconstruction method being used. When dealing with real data this can be tuned for a specific imaging device and correlated with data collection variables such as number of projections and the amount of dose used etc.

4.6 Experimental Results from 3D Data

The previous section showed the effect of an extended field when reconstructing simulated 2D images from 1D projections. This section elaborates the effects of reconstructing 3D densities from 2D data with an an extra region outside the ROI. In the 3D case, the reconstruction problem generally becomes considerably more complicated. For real-data, deconvolution of the point spread function (PSF) is required to obtain accurate reconstructions. Problems due to missing data (missing wedge) are also much more prominent. We used our in-house constrained maximum entropy tomography (COMET) reconstruction package [15] for reconstructing from the 2D TEM data, since it has built-in PSF deconvolution and regularization.

Consider the z-axis to be the direction of the electron beam of the Cryo-Electron Microscope when the data was collected. Let Z_{ROI} be the size of z for which the 3D reconstruction is desired, *i.e.*, the ROI. For the purpose of testing, COMET-based reconstructions can be iterated for volumes increasing in the z-direction, such that $Z_1 < Z_2 < Z_3 < ... < Z_n$, resulting in n-many 3D reconstructions. Each successive reconstruction has an increased volume in the z-direction (Fig. 4.15). Reconstructions in higher volumes are expected to remove more noise from the ROI and should thus have a lower RMS and mean value.

Colloidal silica is made of synthetically manufactured SiO_2 nanoparticles. Like

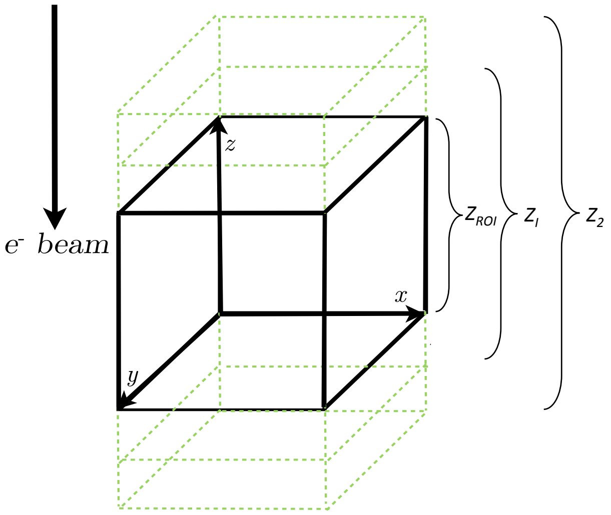
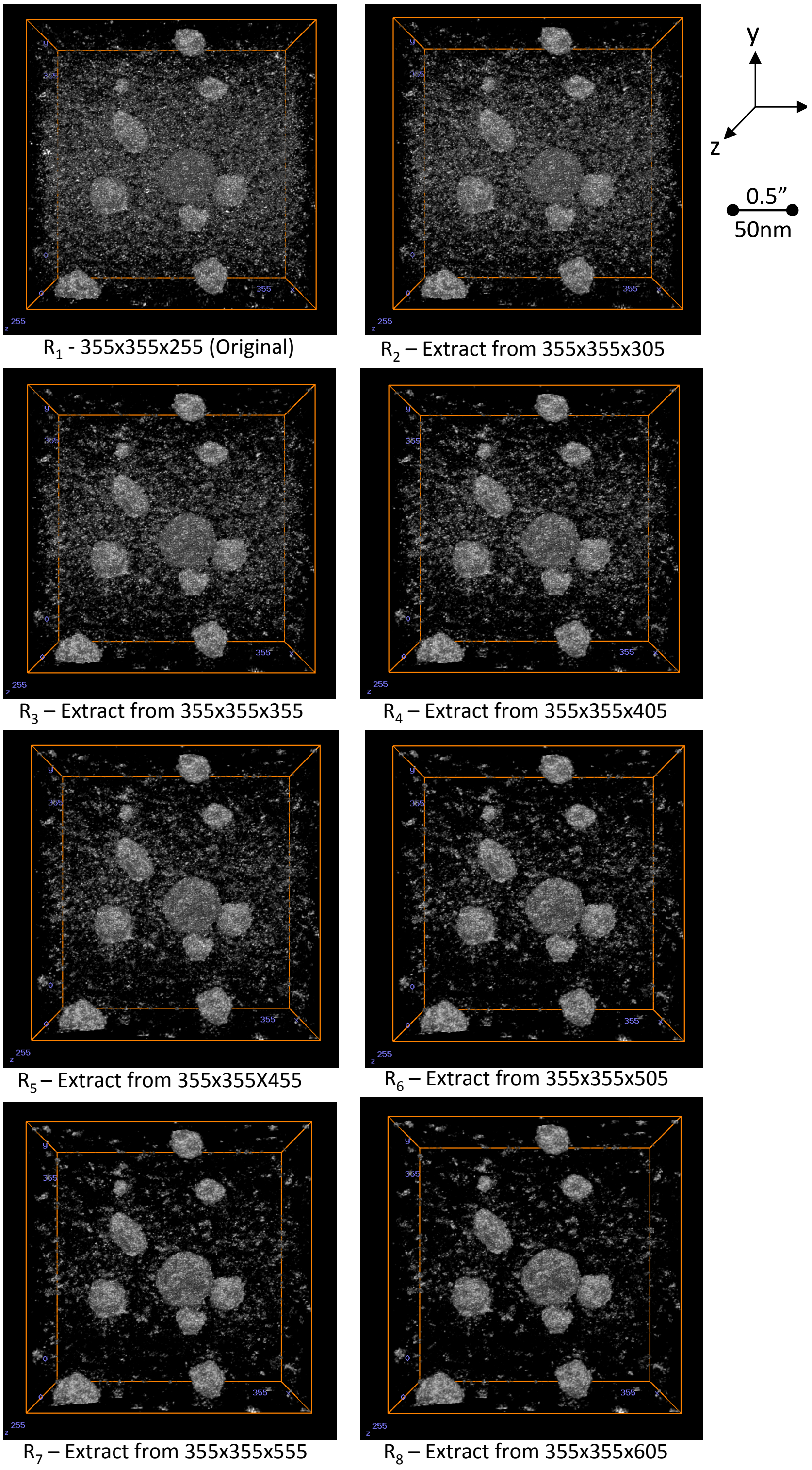


Figure 4.15: Figure showing successive reconstructions on a large volume by increasing in z-direction (*i.e.*, reconstructions with an extended field). The volume may also be increased by increasing the reconstruction space in y-direction.

large proteins, it has a diameter of 25-70 nm and a density of 1.28 g/cm^3, displaying scattering properties comparable to those of proteins [117]. There are several advantages to using colloidal silica as a test object. Since colloidal silica is an inorganic sample it suffers minimally from degradation due to electron beam exposure during data collection. A higher electron dose can be given to such a sample, reducing the shot noise. The colloidal silica sample used here is typically used in paint: BINDZIL of grade 40/130 and was kindly provided by Eka Chemicals, Akzo Nobel, Sweden. Data was collected using a Philips CM200 200keV FEG transmission electron microscope (TEM). The specimen was tilted between ±65° and micrographs were recorded on a CCD detector (F224, TVIPS Gmbh, Germany) every other degree. Eight 3D COMET reconstructions were performed starting with grid points $355 \times 355 \times 255$ (*i.e.*, dimensions in x,y and z axis respectively) to $355 \times 355 \times 605$ with Z increasing incrementally by 50 grid points. Then the common region $355 \times 355 \times 255$ was extracted from all these reconstructions (Fig. 4.16). Images were adjusted to the same contrast threshold. Here we can also clearly see that the background noise decreases over successive reconstructions. Fig. 4.17[2] shows an intensity profile of a single line through a slice of the 3D reconstruction at two different values of Z. It can be clearly seen that at a higher value of Z, the signal is much stronger compared to the noise. Similar results can be achieved by increasing the volume in the y-direction or in both the z- and y-directions. However, the reconstruction space cannot be increased by increasing the x-direction since the sample is tilted around this axis.

[2] Fig. 4.17 was kindly created by Märt Toots using 'R' and is part of our joint publication.



R_1 - 355x355x255 (Original)

R_2 – Extract from 355x355x305

R_3 – Extract from 355x355x355

R_4 – Extract from 355x355x405

R_5 – Extract from 355x355X455

R_6 – Extract from 355x355x505

R_7 – Extract from 355x355x555

R_8 – Extract from 355x355x605

Figure 4.16: Extracted ROIs (355 × 355 × 255) of successive reconstructions of a colloidal silica sample with increasing z. Reconstructions with an extended field (*i.e.* $R_2 - R_8$) have a lower noise within the ROI. The map grid size for one pixel is 5.626Å and the average size of a colloidal silica nano-particle is 31.5nm. The large particle in the center of the reconstruction has lower density as compared to the smaller particles because it is water in the form of crystallized ice. Videos can be viewed in the multimedia supplement with this thesis.

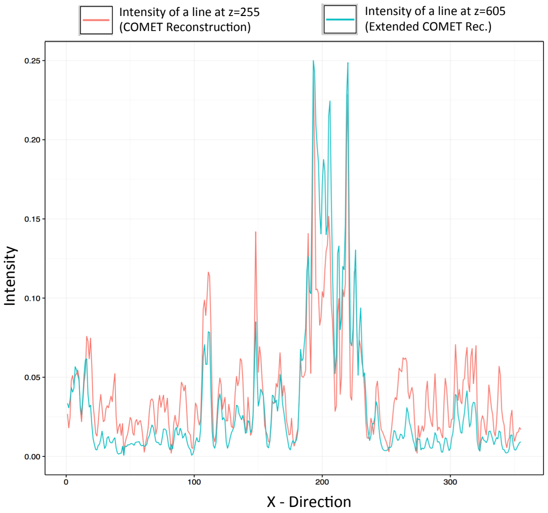


Figure 4.17: Intensity profile of a single line through a slice of the 3D reconstruction at two different values of z. The peak represents the signal. For a reconstruction with an extended field the signal within the ROI is significantly stronger as compared to the noise.

4.7 Limitations and Shortcomings

Limitations of using an extended field can be summarized as follows:

- Extended field does not work with Fourier slice theorem based analytical reconstruction methods such as filtered back projection and its variants. These methods do not allow for redistribution of noise into an extended region. For an analytical method having an extended region outside the ROI would simply mean the same intensity values being distributed along a longer ray path.

- Extended field does not correct for inconsistencies arising from missing data due to limited number of projections or the missing wedge. Although, noise removal due to an extended field might aid other methods which can correct for missing data such as [15].

- The method behaves poorly with weaker regularization methods, better results can be achieved by using more powerful regularization methods as has been shown in proceeding experiments.

4.8 Conclusions

In this chapter we show that extending the field during iterative image reconstruction can render better results when compared to non-extended reconstructions. This principle was tested with extensive simulations using a variety of reconstruction methods. We also propose a heuristic model which explains how this method works and also discusses a limitation of this method. Through extensive empirical simulations we explicitly show that compared to non-extended reconstructions an extended field reconstruction allows the solution of a regularization problem to be much more faithful to the data (projections) by allowing a better fit due to a relatively lower regularization parameter. We also verify the effectiveness of this method by fitting the noise removed and noise added during simulations. We further strengthened our claims by testing the effect using constrained maximum entropy tomography with real cryo-ET data from colloidal silica. The technique is unique in the sense that it can segregate the signal from noise without interfering directly with the signal because it does not involve manipulating individual reconstructed voxels using standard filtering, smoothing, or post-processing procedures, which remove high-frequency details from the reconstruction. Moreover, this method is general and can be extended to be used with other tomographic modalities.

Chapter 5

Algorithm and Architecture Optimization for 2D Discrete Fourier Transforms with Simultaneous Edge Artifact Removal

5.1 Chapter Outline

Two-Dimensional Discrete Fourier Transform (DFT) is a fundamental and computationally intensive algorithm, with a plethora of applications including tomographic reconstructions. 2D images are, in general, non-periodic, but are assumed to be periodic while calculating their DFTs. This leads to cross-shaped artifacts in the frequency domain due to spectral leakage. These artifacts can have critical consequences if the DFTs are being used for further processing, specifically for biomedical applications. In this chapter we present a novel FPGA-based solution to calculate 2D DFTs with simultaneous edge artifact removal for high-performance applications. Standard approaches for removing these artifacts using apodization functions or mirroring, either involve removing critical frequencies or necessitate a surge in computation by significantly increasing the image size. In this chapter a periodic plus smooth decomposition-based approach was used which was optimized to reduce DRAM access and to decrease the number of 1D FFT invocations. 2D FFTs on FPGAs also suffer from the so called 'intermediate storage' or 'memory wall' problem, which is due to limited on-chip memory, increasingly large image sizes and strided column-wise external memory access. A 'tile-hopping' memory mapping scheme is also proposed that significantly improves the bandwidth of the external memory for column-wise reads. The proposed optimizations were tested on a PXIe-based Xilinx Kintex 7 FPGA system communicating with a host PC, which gives the advantage to further expand the design for industrial applications. The proposed high-performance 2D FFT implementation was then used to accelerate filtered back-projection for reconstructing tomographic data. Parts of this work has been published in [118].

5.2 Accelerating Analytical Tomography

Although, this chapter focuses on the architectural development of 2D FFTs with edge artifact removal, the motivation behind this implementation has a deep link to Cryo-ET in specific and tomographic applications in general. As mentioned in Chapter 1 2D FFTs are a fundamental component of the Filtered Back Projection analytical image reconstruction method. The method itself is hailed for its speed and applicability to a wide range of problems. FBP acts as a prior for all methods that have been developed and analyzed in the previous chapters. It is also the cornerstone of AGTV, the method presented in Chapter 3, since the graph prior is constructed on the FBP which enables the method to converge. Without such a prior it would not be possible to use non-local methods for inverse problems. Beyond AGTV, FBP enables the fast convergence of many optimization-based tomography reconstruction methods (often used in a low-pass form to prevent propagation of streaking artifacts).

The ROI Reconstruction Problem in Tomography

Microscopes these days produce large size micrographs or 2D projections usually 4096 × 4096. However, computational constraints don't allow for such large reconstructions to be conducted. Usually the projections are cropped to extract a region of the aligned tomogram which is then reconstructed. This leads to information in the extracted projections which does not belong to the ROI[1]. This erroneous information in the projections can lead to artifacts. If these errors are present in the FBP prior they can be picked up by the iterative method being used for further reconstruction and can be propagated to subsequent iterations. A 4096×4096 FBP reconstruction can remedy this because reconstructing the entire tilt series without extraction will reconstruct the whole tomogram. Then smaller sub-tomograms can be extracted and used as a prior for a more complex optimization-based method. Large size FBP reconstructions are computationally intensive and take several hours.

Why was Fourier-based FBP Chosen Over Analytical Back-projection?

Most practical FBP implementations are done in real space whereby each voxel in the reconstructed density is considered as a sum of the corresponding pixels on the projections. Although, this approach is widely used since it prevents interpolation in the Fourier domain it results in slow reconstructions for large data sizes. There are two major reasons to focus on a 2D FFT-based accelerated FBP FPGA implementation:

- A 2D FFT-based FPGA implementation could be the first step towards implementing a Non-Uniform FFT (NUFFT) [119]. The NUFFT is widely used for tomographic applications since it does not require the interpolation step during FBP reconstructions.
- Large amounts of data rearrangement for real-space FBP can be complicated in

[1] This should not be confused with the ROI problem originating during data collection, reconstructing the entire image will in no way remedy an ROI problem associated with data collection.

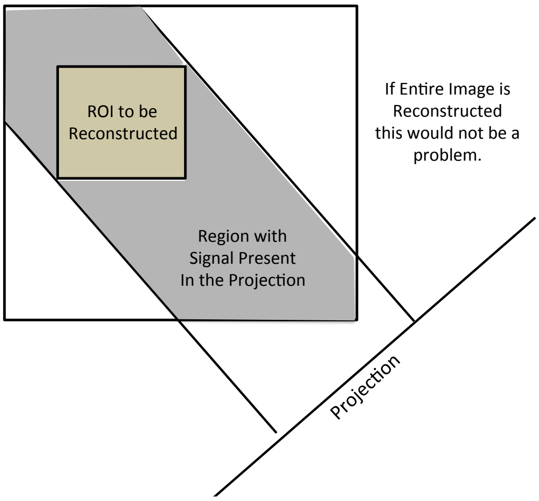

ROI to be Reconstructed
If Entire Image is Reconstructed this would not be a problem.
Region with Signal Present In the Projection
Projection

Figure 5.1: Image showing the ROI reconstruction problem in tomography. Microscopes these days produce large size micrographs (usually 4096 × 4096). However, computational constraints don't allow for such large reconstructions to be conducted. Usually the projections are cropped to extract a region of the aligned tomogram which is then reconstructed. This leads to information in the extracted projections which does not belong to the ROI.

a reconfigurable computing setting, specifically when using high-level reconfigurable computing tools such as LabView FPGA.

Although the motivating factor to implement the 2D FFTs was analytical FBP, deeper analysis of FBP on FPGAs is beyond the scope of this thesis. Here it was only used to check the effectiveness of our 2D FFT implementation explained in the proceeding sections.

5.3 Introduction to 2D FFTs

Discrete Fourier Transform (DFT) is a commonly used and vitally important function for a vast variety of applications including, but not limited to digital communication systems, image processing, computer vision, biomedical imaging and biometrics [120, 121]. Fourier image analysis simplifies computations by converting complex convolution operations in the spatial domain to simple multiplications in the frequency domain. Due to the fundamental nature of 2D DFTs, they are commonly used in a variety of image processing applications such as tomographic image reconstruction [59], non-linear interpolation, texture analysis, tracking, image quality assessment and document analysis [122]. Because of their computational complexity, DFTs often become a computational constraint for applications requiring high throughput and real-time or near real-time operations, specifically for machine vision applications [123]. Image sizes for many of these applications have also increased over the years, further contributing to the problem.

The Cooley-Tukey Fast Fourier Transform (FFT) algorithm [124], first proposed in 1965, reduces the complexity of DFTs from $\mathcal{O}(n^2)$ to $\mathcal{O}(n \log n)$ for a 1D DFT. However, in the case of 2D DFTs, 1D FFTs have to be computed in two-dimensions, increasing the complexity to $\mathcal{O}(n^2 \log n)$, thereby making 2D DFTs a significant bottleneck for real-time machine vision applications [125]. Recently, there has been substantial effort to achieve high-performance implementations of multi-dimensional FFTs to overcome this constraint [118, 123, 125–131]. Due to their inherent parallelism and reconfigurability Field Programmable Gate Arrays (FPGAs) are attractive targets for accelerating FFT computations. Being a highly flexible platform FPGAs can fully exploit the parallel nature of the FFT algorithm. 2D FFTs are generally calculated in stages where all elements of the first stage must be available before the second stage can be calculated. This creates the so-called *'intermediate storage'* problem associated with strided external memory access, specifically for large datasets.

While calculating 2D DFTs it is assumed that the image is periodic, which is usually not the case. The non-periodic nature of the image leads to artifacts in the Fourier transform, usually known as edge artifacts or series termination errors. These artifacts appear as several crosses of high-amplitude coefficients in the frequency domain, as seen in [132, 133]. Such edge artifacts can be passed to subsequent stages of processing and in biomedical applications they may lead to critical misinterpretations of results. Efficiently removing such artifacts without compromising resolution is a major problem. Moreover, simultaneously removing these spurious artifacts while calculating the 2D FFT adds to the existing complexity of the 2D FFT kernel.

5.3.1 Contributions

In this chapter we present solutions for a high-performance 2D DFT with simultaneous edge artifact removal for applications which require high frame rate 2D FFTs such as real-time medical imaging systems, machine vision for control etc. This work builds on our previous work presented in [118]. Major contributions include:

- We propose an algorithmic optimization for Periodic plus Smooth Decomposition (PSD) for edge artifact removal presented in [122].

- Based on our optimized periodic plus smooth decomposition (OPSD) we propose an architecture that can reduce the access to DRAM and can decrease the number of 1D FFT invocations by performing column-by-column operations on the fly [118].

- Since OPSD is heavily dependent on an efficient FPGA-based 2D FFT implementation which is limited by DRAM access problems we design a memory mapping scheme which can reduce row activation overhead while accessing columns of data from the DRAM.

- We use our implementation as an accelerator for filtered back-projection (FBP), an analytical tomographic reconstruction method, and show that for large datasets our 2D FFT with edge artifact removal can significantly improve reconstruction run time.

The chapter follows FPGA image processing design methodology outlined in [133, 134], which involves carefully profiling the software solution to understand computational bottlenecks and overcoming them through careful reformulation of the algorithm within a parallel hardware framework.

5.4 Background

5.4.1 High Performance 2D FFTs using FPGAs

There are several resource-efficient, high-throughput implementations of multidimensional DFTs on a variety of different platforms. Many of these implementations are software-based and have been optimized for efficient performance on general-purpose processors (GPPs), for example Intel MKL [128], FFTW [126] and Spiral [127]. Implementations on GPPs can be readily adapted for a variety of scenarios. However, GPPs consume more power and are not ideal for real-time embedded applications. Several Application-Specific Integrated Circuit (ASIC)-based implementations have also been proposed [135–137], but since it is not easy to modify ASIC implementations, they are not cost-effective solutions for rapid prototyping of image processing systems. GPUs on the other hand can achieve relatively high throughput but are energy inefficient. Moreover, they often have limited bandwidth and limit the mobility of large scale imaging systems.

Due to their inherent parallelism and reconfigurability, FPGAs are attractive for accelerating FFT computations, since they fully exploit the parallel nature of the FFT algorithm. FPGAs are particularly an attractive target for medical and biomedical imaging apparatus and instruments such as electron microscopes and tomographic scanners. Such devices don't have to be manufactured in bulk to justify application specific solutions, and require high bandwidth. Moreover, increasing mobility and portability is a future objective for many medical imaging systems. FPGAs are also more efficient for prototyping machine vision applications since they are relatively more fine-grained when compared to GPPs and GPUs and can serve as a bridge between a general purpose and application specific acceleration solutions.

5.4.2 DRAM Intermediate Storage Problem

There have been several high-throughput 2D FFT FPGA-based implementations over the past few years. Most of these rely on repeated invocations of 1D FFTs by row and column decomposition (RCD) with efficient use of memory [123, 125, 129, 138, 139]. RCD makes use of the fact that a 2D Fourier transform is separable and can be implemented in stages, *i.e.*, a row-by-row 1D FFT can be proceeded by a column-by-column 1D FFT with intermediate storage (Fig. 5.2). Most of the previous RCD-based 2D FFT FPGA implementations have two major design challenges: 1) The 1D FFT implementation needs to have a reasonably high-throughput and needs to be resource-efficient. Moreover, spatial parallelism needs to be exploited by running several 1D FFTs simultaneously. 2) External DRAM needs to be efficiently addressed and to have a high bandwidth.

Since the column-by-column 1D FFT requires data from all rows intermediate storage becomes a major problem for large datasets. Many implementations rely on local

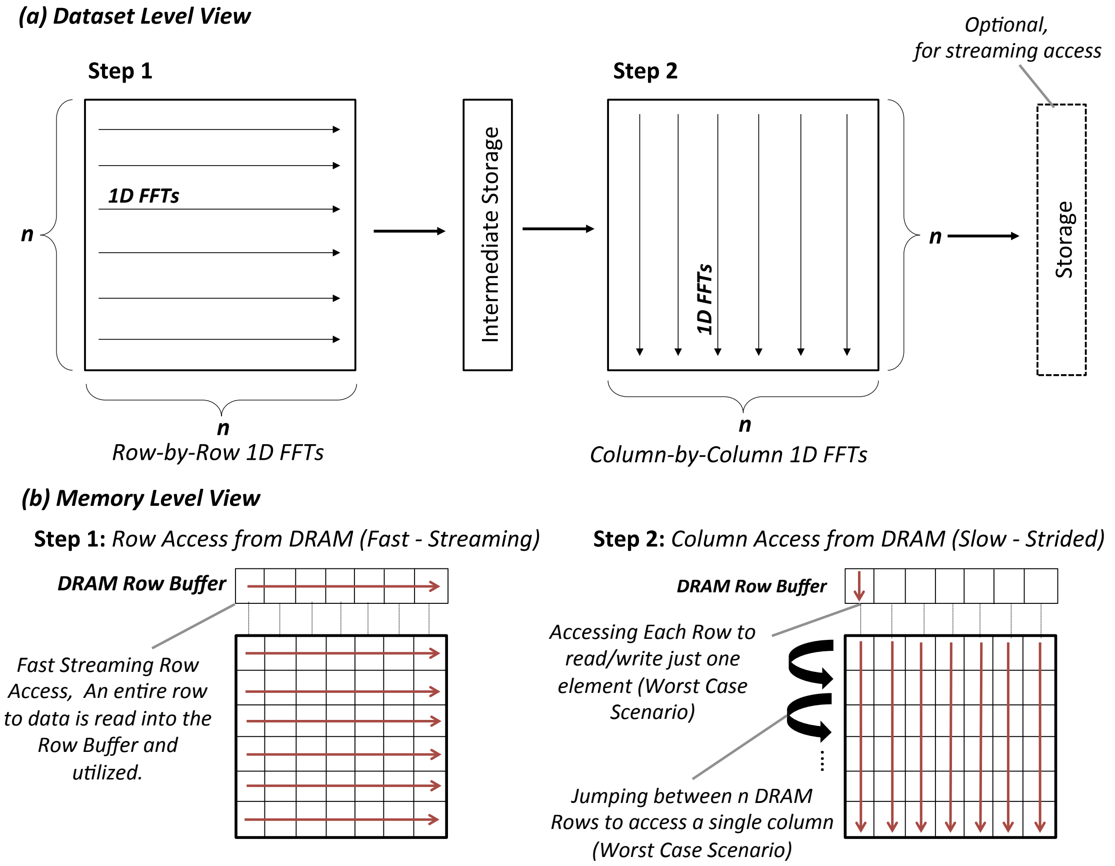

(a) Dataset Level View
Step 1
1D FFTs
n
n
Row-by-Row 1D FFTs
Intermediate Storage
Step 2
1D FFTs
n
n
Column-by-Column 1D FFTs
Optional,
for streaming access
Storage
(b) Memory Level View
Step 1: Row Access from DRAM (Fast - Streaming)
DRAM Row Buffer
Fast Streaming Row Access, An entire row to data is read into the Row Buffer and utilized.
Step 2: Column Access from DRAM (Slow - Strided)
DRAM Row Buffer
Accessing Each Row to read/write just one element (Worst Case Scenario)
Jumping between n DRAM Rows to access a single column (Worst Case Scenario)

Figure 5.2: a) An overview of Row-Column Decomposition (RCD) for 2D FFT implementation. Intermediate Storage is required because all elements of the row-by-row operations must be available for column-by-column processing. b) An overview of strided Column-wise access from DRAM as compared to trivial row-wise access. An entire row of elements must be read into the row buffer even to access a single element within a specific row.

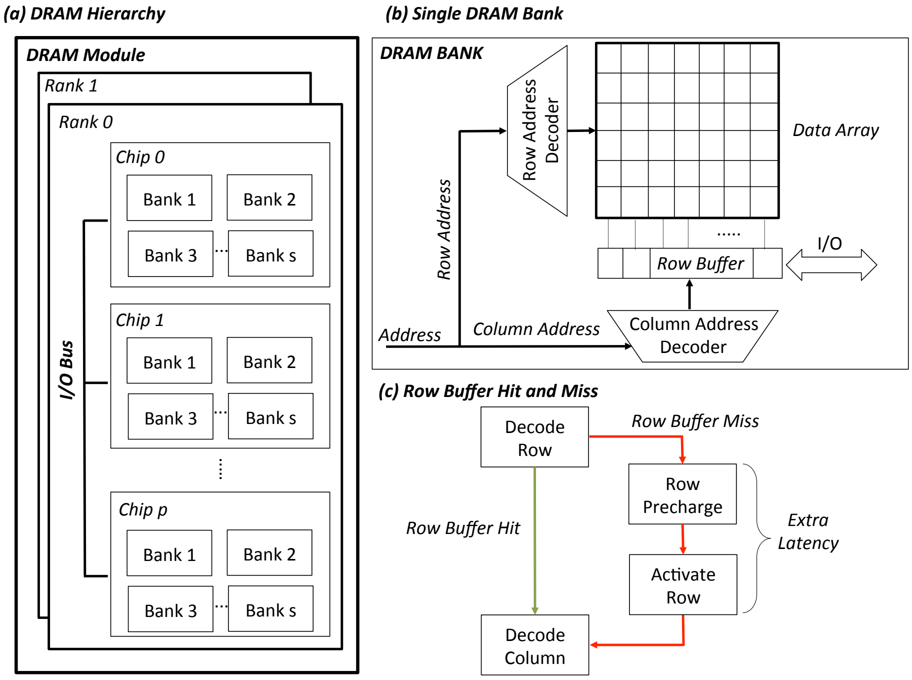


Figure 5.3: a) An overview of the DRAM hierarchy. b) Image showing the structure of a single DRAM bank. c) Flow chart explaining additional latency introduced when a new row has to be referred into the row buffer to access a specific element.

memory such as resource-implemented Block RAM for intermediate storage which is not possible for large datasets [138]. Large datasets have to be offloaded to external DRAM because only a portion of the dataset that fits on the chip can be operated on at a given time. For complex image processing applications this means repeated storage and access to the external memory during every stage of processing.

As shown in Fig. 5.3a DRAM hierarchy from top to bottom is: rank, chip, bank, row and column. Each DRAM bank (Fig. 5.3b) has a *row buffer* that holds the most recently referred row. There is only one *row buffer* per bank which means only one row from the data-grid can be accessed at once. Before accessing a row it has to be activated by transferring the contents from internal capacitor storage into a set of parallel sense amplifiers. The *row buffer* is the so-called *'fast buffer'*, because when a row is activated and placed in the buffer, any element can be accessed at random.

If a new row has to be activated and accessed into the row buffer a *row buffer miss* occurs and requires a higher latency, A_{miss} (Fig. 5.3c). On the contrary if the desired row is already in the buffer a *row buffer hit or page hit* occurs and the latency to access elements is substantially lower, A_{hit}. This implies $A_{miss} = A_{hit} + C_r$, where C_r is the overhead associated with accessing a new row to read a specific element (Fig. 2c) [129]. There is also overhead involved in writing the row back to the data-grid (pre-charge), say C_w. However, both C_r and C_w can be concealed by interleaving (smart switching between banks). Since row-wise access is trivial the row-by-row 1D FFT part of RCD-based 2D FFT is simple. However, once the row-by-row 1D FFT is stored in the DRAM in standard row-major order, to access a single column, each row of the DRAM has to be accessed into the row buffer rendering the read process extremely inefficient. This is typically the major bottle neck for high-throughput 2D FFTs (Fig. 5.2b). We address this problem by designing a custom memory mapping scheme (Section V).

5.4.3 Edge Artifacts

While calculating 2D DFTs it is assumed that the image is periodic, which is usually not the case. The non-periodic nature of the image leads to artifacts in the Fourier transform, usually known as edge artifacts or series termination errors. These artifacts appear as several crosses of high-amplitude coefficients in the frequency domain (Fig. 5.4b). Such edge artifacts can be passed to subsequent stages of processing and in biomedical applications they may lead to critical misinterpretations of results. No current 2D FFT FPGA implementation addresses this problem directly. These artifacts may be removed during pre-processing, using mirroring, windowing, zero padding or post-processing, *e.g.,* filtering techniques. These techniques are usually computationally intensive, involve an increase in image size, and also tend to modify the transform.

The most common approach is by ramping the image at corner pixels to slowly attenuate the edges. Ramping is usually accomplished by an apodization function such as a Tukey (tapered cosine) or a Hamming window, which smoothly reduces the intensity to zero. Such an approach can be implemented on an FPGA as a pre-processing operation by storing the window function in a Look-up Table (LUT) and multiplying it with the image stream before calculating the FFT [133]. Although this approach is not extremely computationally intensive for small images, it inadvertently removes necessary information from the image. Loss of this information may have serious consequences if the image is being further processed with several other images to reconstruct a final image that is used for diagnostics or other decision-critical applications. Another common method is by mirroring the image from $N \times N$ to $2N \times 2N$. Doing so makes the image periodic, and reduces edge artifacts. However, this not only increases the size of the image by 4×, but also makes the transform symmetric, which generates an inaccurate phase component.

Simultaneously removing the edge artifacts while calculating a 2D FFT imposes an additional design challenge, regardless of the method used. However, these artifacts must be removed in applications where they may be propagated to subsequent processing levels. An ideal method for removing these artifacts should involve making the image periodic while removing minimal information from the image. PSD, first presented by Moisan [122] and used in [140–142], is an ideal method for removing edge artifacts (specifically for biomedical applications) because it does not directly intervene with pixels beside those of the boundary and does not increase image size. Moreover, its inherently parallel nature makes it ideal for a high-throughput, FPGA-based implementation. We have further optimized the original PSD decomposition algorithm to make the overall implementation much more efficient, by decreasing the number of required 1D FFT invocations and by reducing external DRAM utilization (Section IV).

5.4.4 LabView FPGA Semi-High Level Design Environment

A major concern while designing complex image processing hardware accelerators is how to fully harness the divide-and-conquer approach. Algorithms that have to be mapped to multiple FPGAs are often marred by communication problems, and custom FPGA boards reduce flexibility for large-scale and evolving designs. For rapid-prototyping of our algorithms we used LabView FPGA 2016 (National Instruments), a robust data-flow-based graphical design environment. LabView FPGA provides integration with National Instruments (NI) Xilinx-based reconfigurable hardware, allowing efficient communication with a host PC and high-throughput communication between multiple FPGAs through a PXIe (PCI eXtentions for Industry Express) bus. LabView FPGA also enables us to integrate external Hardware Description Language (HDL) code and gives us the flexibility to expand our design for future processing stages. We used NI PXIe-7976R FPGA boards that have a Xilinx Kintex 7 FPGA and 2GB high-bandwidth (10GB/s) external memory. This platform has already been extensively used for rapid-prototyping of communication standards and protocols before moving to ASIC designs. The optimizations and designs we present here are scalable to any reconfigurable computing-based system.

5.5 Periodic Plus Smooth Decomposition (PSD) for Edge Artifact Removal

Periodic plus smooth decomposition (PSD) involves decomposing the image into a periodic and smooth component to remove edge artifacts with minimal loss of information from the image [122]. This section presents an overview of the PSD algorithm and profiles the algorithm for possible parallelization and optimization to achieve an efficient FPGA implementation.

Let us have discrete n by m gray-scale image $\boldsymbol{I}$ on a finite domain $\Omega = \{0, 1, \ldots, n-1\} \times \{0, 1, \ldots, m-1\}$. The discrete Fourier transform (DFT) of $\boldsymbol{I}$ is defined as

$$\hat{\boldsymbol{I}}(s,t) = \sum_{(i,j)\in\Omega} \boldsymbol{I}(i,j) \exp\left(-\iota 2\pi \left(\frac{si}{n} + \frac{tj}{m}\right)\right) \tag{5.1}$$

This is equivalent to a matrix multiplication $\boldsymbol{WIV}$, where

$$\boldsymbol{W} = \begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & w & w^2 & \ldots & w^{n-1} \\ 1 & w^2 & w^4 & \ldots & w^{2(n-1)} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & w^{n-2} & w^{2(n-2)} & \ldots & w^{(n-2)(n-1)} \\ 1 & w^{n-1} & w^{2(n-1)} & \ldots & w^{(n-1)(n-1)} \end{pmatrix} \tag{5.2}$$

and

$$w^k = \exp\left(-\iota\frac{2\pi}{n}\right)^k = \exp\left(-\iota\frac{2\pi k}{n}\right). \tag{5.3}$$

Algorithm 5 Periodic Plus Smooth Decomposition (PSD)

```
INPUT: $\boldsymbol{I}(i,j)$ of size $n \times m$
Step A: Compute the 2D DFT of image $\boldsymbol{I}(i,j)$ :
$\boldsymbol{I}(i,j) \xrightarrow{\mathscr{F}} \hat{\boldsymbol{I}}(s,t)$
Step B: Compute periodic border $\boldsymbol{B}$ :
if $(i = 0 \vee i = n-1)$ then
    $\boldsymbol{R}(i,j) \leftarrow \boldsymbol{I}(n-1-i,j) - \boldsymbol{I}(i,j)$
else
    $\boldsymbol{R}(i,j) \leftarrow 0$
end if
if $(j = 0 \vee j = m-1)$ then
    $\boldsymbol{C}(i,j) \leftarrow \boldsymbol{I}(i,m-1-j) - \boldsymbol{I}(i,j)$
else
    $\boldsymbol{C}(i,j) \leftarrow 0$
end if
$\boldsymbol{B} \leftarrow \boldsymbol{R} + \boldsymbol{C}$
Step C: Compute the 2D DFT of $\hat{\boldsymbol{B}}(i,j)$ :
$\boldsymbol{B}(i,j) \xrightarrow{\mathscr{F}} \hat{\boldsymbol{B}}(s,t)$
Step D: Compute the Smooth Component $\hat{\boldsymbol{S}}(s,t)$ :
$\hat{\boldsymbol{D}}(s,t) \leftarrow 2\cos\frac{2\pi s}{n} + 2\cos\frac{2\pi t}{m} - 4$
$\hat{\boldsymbol{S}}(s,t) \leftarrow \hat{\boldsymbol{B}}(s,t) \div \hat{\boldsymbol{D}}(s,t)$
Step E: Compute the Periodic Component $\hat{\boldsymbol{P}}(s,t)$ :
$\hat{\boldsymbol{P}}(s,t) \leftarrow \hat{\boldsymbol{I}}(s,t) - \hat{\boldsymbol{S}}(s,t)$
OUTPUT: $\hat{\boldsymbol{P}}(s,t)$, $\hat{\boldsymbol{S}}(s,t)$
```

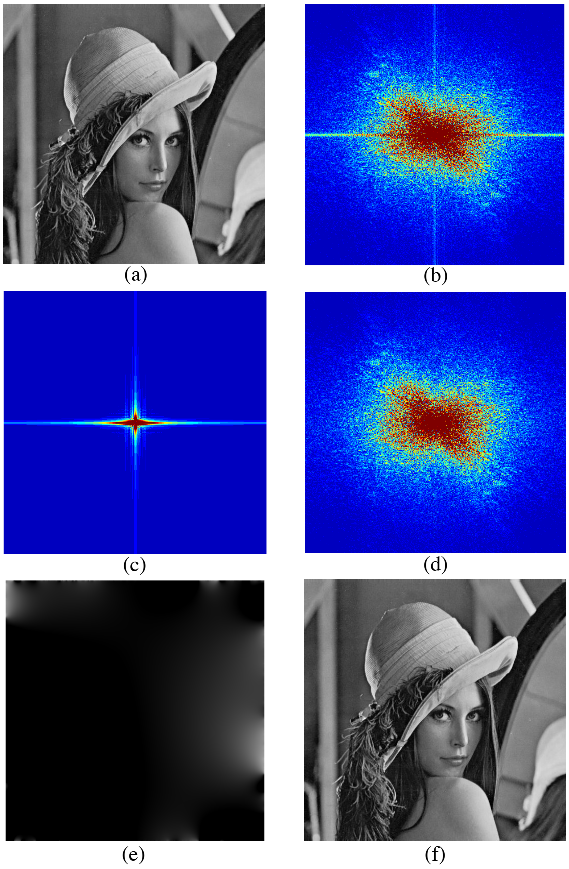


Figure 5.4: 1a) An image with non-periodic boundary. 1b) 2D DFT of 1a. 1c) DFT of the Smooth Component ,*i.e.*, the removed artifacts from 1a. 1d) Periodic Component i.e DFT of 1a with Edge Artifacts removed. 1e) Reconstructed Smooth Component. 1f) Reconstructed Periodic Component.

$\boldsymbol{V}$ has the same structure as $\boldsymbol{W}$ but is m-dimensional. Since w^k has period n which means that $w^k = w^{k+ln}$, $\forall k, l \in \mathbb{N}$ and therefore,

$$\boldsymbol{W} = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 & 1 \\ 1 & w & w^2 & \dots & w^{n-2} & w^{n-1} \\ 1 & w^2 & w^4 & \dots & w^{n-4} & w^{n-2} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & w^{n-2} & w^{n-4} & \dots & w^4 & w^2 \\ 1 & w^{n-1} & w^{n-2} & \dots & w^2 & w^1 \end{pmatrix} \tag{5.4}$$

Since in general $\boldsymbol{I}$ is not (n, m)-periodic, there will be high amplitude edge artifacts present in the DFT stemming from sharp discontinuities between the opposing edges of the image as shown in Fig. 5.4b. [122] proposed a decomposition of I into a periodic component $\boldsymbol{P}$, that is periodic and captures the essence of the image with all high frequency details, and a smoothly varying background $\boldsymbol{S}$, that recreates the discontinuities at the borders. So, $\boldsymbol{I} = \boldsymbol{P} + \boldsymbol{S}$. Periodic plus smooth decomposition can be computed by first constructing a border image $\boldsymbol{B} = \boldsymbol{R} + \boldsymbol{C}$, where $\boldsymbol{R}$ represents the boundary discontinuities when transitioning row-wise and $\boldsymbol{C}$ when going column-wise,

$$\begin{aligned} \boldsymbol{R}(i,j) &= \begin{cases} \boldsymbol{I}(n-1-i,j) - \boldsymbol{I}(i,j), & i = 0 \text{ or } i = n-1 \\ \mathbf{0}, & \text{otherwise} \end{cases} \\ \boldsymbol{C}(i,j) &= \begin{cases} \boldsymbol{I}(i,m-1-j) - \boldsymbol{I}(i,j), & j = 0 \text{ or } j = m-1 \\ \mathbf{0}, & \text{otherwise} \end{cases} \end{aligned} \tag{5.5}$$

It is obvious that the structure of the border image $\boldsymbol{B}$ is simple with nonzero values only in the edges as shown below:

$$\boldsymbol{B} = \boldsymbol{R} + \boldsymbol{C} = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1,m-1} & b_{1m} \\ b_{21} & 0 & \dots & 0 & -b_{21} \\ \dots & \dots & \dots & \dots & \dots \\ b_{n-1,1} & 0 & \dots & 0 & -b_{n-1,1} \\ b_{n1} & -b_{12} & \dots & -b_{1,m-1} & -b_{nm} \end{pmatrix}. \tag{5.6}$$

The DFT of the smooth component $\boldsymbol{S}$ can be then found by the following formula:

$$\hat{\boldsymbol{S}}(s,t) = \frac{\hat{\boldsymbol{B}}(s,t)}{2\cos\frac{2\pi s}{n} + 2\cos\frac{2\pi t}{m} - 4}, \quad \forall (s,t) \in \Omega \backslash \{(0,0)\}. \tag{5.7}$$

The DFT of the image $\boldsymbol{I}$ with edge artifacts removed is then $\hat{\boldsymbol{P}} = \hat{\boldsymbol{I}} - \hat{\boldsymbol{S}}$. Fig. 5.4c and 5.4d show the DFT of the smooth and periodic components, respectively. Fig. 5.4e and Fig. 5.4f show the reconstructed periodic and smooth components. On reconstruction, it is evident that there is negligible visual difference between the actual image and the periodic reconstructed image.

5.5.1 Profiling PSD for FPGA Implementation

Algorithm 5 summarizes the overall PSD implementation. There are several ways of arranging the algorithm. We have arranged it so that DFTs of the periodic and smooth components are readily available for further processing stages. For best results, both the periodic and smooth components should undergo similar processing stages and should be added back together before displaying the result. However, depending on the application it might be acceptable to discard the smooth component completely. For a $n \times m$ image step A and C have a complexity of $\mathcal{O}(nm\log(nm))$ and steps B and D have complexity $\mathcal{O}(m+n)$ and $\mathcal{O}(mn)$ respectively. Computationally the performance of PSD is limited by step A and C. Fig. 5.5 shows a proposed top-level architecture where step A and steps B, C and D are completed on separate FPGAs while step E can be done on the host PC. The overall performance may be limited by FPGA 2 where most of the serial part of the algorithm lies. There are two major factors which limit the throughput of such a design:

1. While FPGA 1 and FPGA 2 can run in parallel the result of step A from FPGA 1 has to be stored on the host PC while steps B, C and D are completed on FPGA 2 before step E can be completed on the host PC.

2. The DRAM intermediate storage problem explained in section 5.4.2 and Fig.

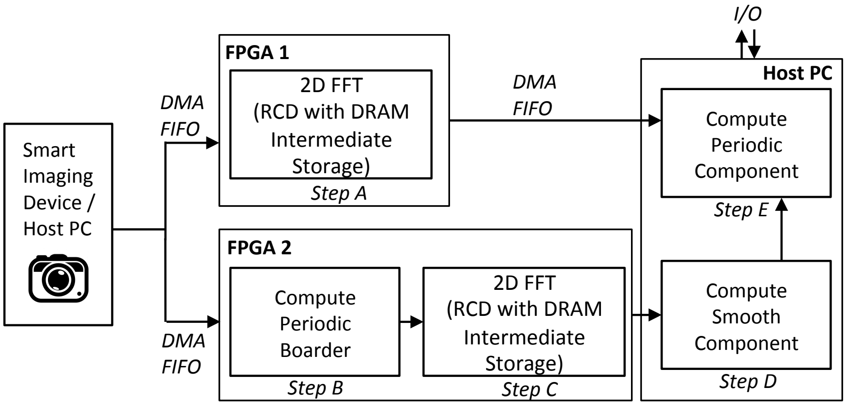

I/O
Host PC
FPGA 1
2D FFT (RCD with DRAM Intermediate Storage)
Step A
DMA FIFO
DMA FIFO
Smart Imaging Device / Host PC
Compute Periodic Component
Step E
FPGA 2
Compute Periodic Boarder
Step B
2D FFT (RCD with DRAM Intermediate Storage)
Step C
DMA FIFO
Compute Smooth Component
Step D

Figure 5.5: A top-level architecture for OPSD using two FPGAs and a host PC connected over a high-bandwidth bus. The steps are associated with algorithm 5.

5.4.2 has to be addressed since strided access to the DRAM for column-wise operations can significantly limit throughput.

As for (1) it has been addressed in the next section where we make use of the inherent symmetry of the boundary image to reduce the time required to compute the 2D FFT of the boundary image. As for (2) it has been addressed by designing a semi-custom memory mapping controller which *"tiles"* the DRAM floor and *"hops"* between several tiles so as to minimize strided memory access.

5.6 Optimized Periodic Plus Smooth Decomposition (OPSD)

In this section we optimize [2] the original PSD algorithm so that it can be effectively configured on an FPGA by reducing the number of 1D FFT invocations, reducing DRAM access and eliminating column-wise 1D FFT operations.

On inspecting Equation (6) we realize that the boundary image $\boldsymbol{B}$ is symmetrical in the sense that boundary rows and columns are an algebraic negation of each other. In total $\boldsymbol{B}$ has $n+m-1$ unique elements, with the following relations between corners with respect to columns and rows:

$$\begin{aligned} b_{11} &= r_{11} + c_{11} \\ b_{1m} &= r_{1m} - c_{11} \\ b_{n1} &= -r_{11} + c_{n1} \\ b_{nm} &= -r_{1m} - c_{n1} \end{aligned} \tag{5.8}$$

$$\implies b_{nm} = -b_{11} - b_{1m} - b_{n1} \tag{5.9}$$

[2] The optimization presented in Equations 5.8-5.16 was suggested and tested on software by Märt Toots and is part of our joint publication [118].

In computing the FFT of $\boldsymbol{B}$ one normally proceeds by first running 1D FFTs column-by-column and then 1D FFTs row-by-row or vice versa. An FFT of a column vector $\boldsymbol{v}$ with length n is $\boldsymbol{Wv}$, where $\boldsymbol{W}$ is given in eq. (5.4). The column-wise FFT of the matrix $\boldsymbol{B}$ is then

$$\hat{\boldsymbol{B}} = \boldsymbol{WB}. \tag{5.10}$$

Let us have a closer look on the first column, denoted by $\boldsymbol{B_{\cdot 1}}$. The 1D FFT of this vector is

$$\begin{aligned}\hat{\boldsymbol{B}}_{\boldsymbol{\cdot 1}} = \boldsymbol{WB_{\cdot 1}} = &\begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & w & \dots & w^{n-1} \\ 1 & w^2 & \dots & w^{2(n-1)} \\ \dots & \dots & \dots & \dots \\ 1 & w^{n-2} & \dots & w^{(n-2)(n-1)} \\ 1 & w^{n-1} & \dots & w^{(n-1)(n-1)} \end{pmatrix} \begin{pmatrix} b_{11} \\ b_{21} \\ b_{31} \\ \dots \\ b_{n-1,1} \\ b_{n1} \end{pmatrix} \\ = &\begin{pmatrix} \sum_{i=1}^n b_{i1} \\ \sum_{i=1}^n b_{i1} w^{i-1} \\ \sum_{i=1}^n b_{i1} w^{2(i-1)} \\ \dots \\ \sum_{i=1}^n b_{i1} w^{(n-2)(i-1)} \\ \sum_{i=1}^n b_{i1} w^{(n-1)(i-1)} \end{pmatrix}\end{aligned} \tag{5.11}$$

It can be shown that the 1D FFT of the column $j \in \{2, 3, \dots, m-1\}$ is

$$\begin{aligned}\hat{\boldsymbol{B}}_{\boldsymbol{\cdot j}} = \boldsymbol{WB_{\cdot j}} = &\begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & w & \dots & w^{n-1} \\ 1 & w^2 & \dots & w^{2(n-1)} \\ \dots & \dots & \dots & \dots \\ 1 & w^{n-2} & \dots & w^{(n-2)(n-1)} \\ 1 & w^{n-1} & \dots & w^{(n-1)(n-1)} \end{pmatrix} \begin{pmatrix} b_{1j} \\ 0 \\ 0 \\ \dots \\ 0 \\ -b_{1j} \end{pmatrix} \\ =&b_{1j} \begin{pmatrix} 0 \\ 1-w^{n-1} \\ 1-w^{n-2} \\ \dots \\ 1-w^2 \\ 1-w \end{pmatrix} = b_{1j}\boldsymbol{\nu}\,,\end{aligned} \tag{5.12}$$

The 1D FFT of the last column $\boldsymbol{B_{\cdot m}}$ is

$$\hat{\boldsymbol{B}}_{\boldsymbol{\cdot m}} = \boldsymbol{WB_{\cdot m}} \tag{5.13}$$

Table 5.1: Comparing Mirroring, PSD and OPSD

Algorithm	DRAM Access	DFT
	Points	Points
Mirroring	$8NM$	$8NM$
P+S Decomposition (PSD)	$4NM$	$4NM$
Optimized PSD (Proposed)	$3NM + N + M - 1$	$3NM + M$

$$= \begin{pmatrix} -\sum_{i=1}^{n} b_{i1} \\ -\sum_{i=1}^{n} b_{i1} w^{i-1} + (b_{11} + b_{1m})(1 - w^{n-1}) \\ -\sum_{i=1}^{n} b_{i1} w^{2(i-1)} + (b_{11} + b_{1m})(1 - w^{2(n-1)}) \\ \dots \\ -\sum_{i=1}^{n} b_{i1} w^{(n-2)(i-1)} + (b_{11} + b_{1m})(1 - w^{(n-2)(n-1)}) \\ -\sum_{i=1}^{n} b_{i1} w^{(n-1)(i-1)} + (b_{11} + b_{1m})(1 - w^{(n-1)(n-1)}) \end{pmatrix} \tag{5.14}$$

$$\hat{\boldsymbol{B}}_{\cdot \boldsymbol{m}} = -\,\hat{\boldsymbol{B}}_{\cdot \boldsymbol{1}} + (b_{11} + b_{1m})\boldsymbol{\nu}\,. \tag{5.15}$$

So, the column-wise FFT of the matrix $\boldsymbol{B}$ is

$$\hat{\boldsymbol{B}} = \begin{pmatrix} \hat{\boldsymbol{B}}_{\cdot \boldsymbol{1}} & b_{12}\boldsymbol{\nu} & \dots & b_{1,m-1}\boldsymbol{\nu} & -\hat{\boldsymbol{B}}_{\cdot \boldsymbol{1}} + (b_{11} + b_{1m})\boldsymbol{\nu} \end{pmatrix}. \tag{5.16}$$

To compute the column-by-column 1D FFT of the matrix, $\boldsymbol{B}$, we only have to compute the FFT of the first vector and then use the appropriately scaled vector, $\boldsymbol{\nu}$, to derive the remainder of the columns. The row-by-row FFT has to be calculated in a row burst normal way. Algorithm 6 presents a summary of the shortcut for calculating $\hat{\boldsymbol{B}}(s, t)$. The steps presented in Algorithm 6 can replace parts of Algorithm 5. By reducing column-by-column 1D FFT computations for the boundary image, this method can significantly reduce the number of 1D FFT invocations, reduce the overall DRAM access and eliminate problematic column-wise strided DRAM access for an efficient FPGA-based implementation. For column-wise operations a single 1D FFT of size m is required rather than nm 1D FFTs of size m. Moreover, since one has to simply store one column of data it can be stored on the on-chip local memory (BRAM or SRAM). This can be implemented by temporarily storing the initial vector $\hat{\boldsymbol{B}}_{\cdot \boldsymbol{1}}$ and scaling factors b_{1j} in the block RAM/register memory, drastically reducing DRAM access and lowering the number of required 1D FFT invocations. Performance evaluation for this has been presented in the results section.

Table 5.1 shows a comparison of Mirroring, PSD and our proposed OPSD with respect to DRAM access points. Mirroring has been used for comparison purposes because it is an alternative technique that reduces edge artifacts while maintaining maximum amplitude information. However, due to replication of the image most of

Algorithm 6 Proposed Symmetrically Optimized Computation of $\hat{\boldsymbol{B}}(s,t)$

```
INPUT: $\boldsymbol{B}(i,j)$ of size $n \times m$
$\textit{2D } \mathscr{F}(\boldsymbol{B}) \Leftrightarrow \boldsymbol{B} \xrightarrow{\mathscr{F}_{cw}} \hat{\boldsymbol{B}}_{\boldsymbol{cw}} \xrightarrow{\mathscr{F}_{rw}} \hat{\boldsymbol{B}}$
Column-by-Column DFT via Symmetrical Short-cut:
$\boldsymbol{B}_{\cdot\boldsymbol{1}} \xrightarrow{\mathscr{F}} \hat{\boldsymbol{B}}_{\cdot\boldsymbol{1}}$
while $1 < j < m$ do
    $\hat{\boldsymbol{B}}_{\cdot\boldsymbol{j}} \leftarrow b_{1j}\boldsymbol{\nu}$
end while
$\hat{\boldsymbol{B}}_{\cdot\boldsymbol{m}} \leftarrow -\hat{\boldsymbol{B}}_{\cdot\boldsymbol{1}} + (b_{11} + b_{1m})\boldsymbol{\nu}$
$\hat{\boldsymbol{B}}_{\boldsymbol{cw}} \leftarrow \textit{Concatenate } \hat{\boldsymbol{B}}_{\cdot\boldsymbol{1}}\ \hat{\boldsymbol{B}}_{\cdot\boldsymbol{2}}\ \ldots\ \hat{\boldsymbol{B}}_{\cdot\boldsymbol{m-1}}\ \hat{\boldsymbol{B}}_{\cdot\boldsymbol{m}}$

Row-by-Row DFT:
$\hat{\boldsymbol{B}}_{\boldsymbol{cw}} \xrightarrow{\mathscr{F}_{rw}} \hat{\boldsymbol{B}}(s,t)$
OUTPUT: $\hat{\boldsymbol{B}}(s,t)$
cw: column-wise/column-by-column
rw: row-wise/row-by-row
```

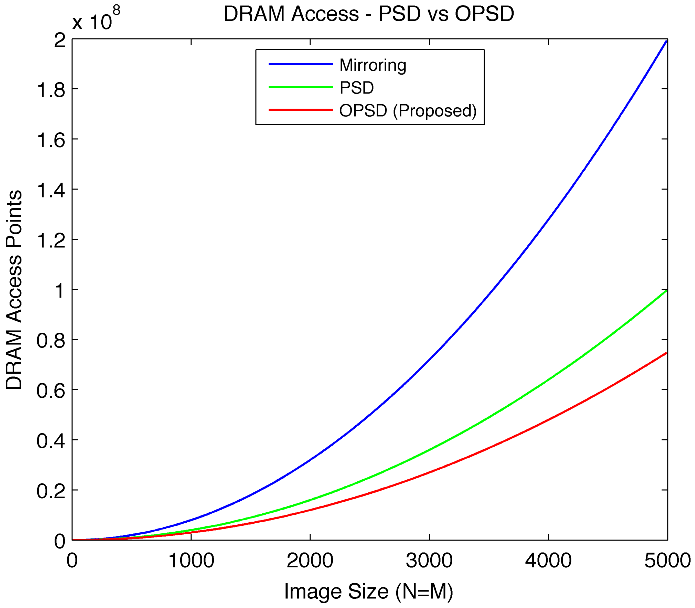


Figure 5.6: Graph showing DRAM access (equal to number of DFT points to be computed) with increasing image size for Mirroring, Periodic Plus Smooth Decomposition (PSD) and our proposed Optimized Period Plus Smooth Decomposition (OPSD).

the phase information is lost. Fig. 5.6 graphically shows that our OPSD method significantly reduces reading from external memory and reduces the overall number of DFT computations required. It should be noted that such optimization is only possible for either column-wise or row-wise operations because after either of these operations the output is not symmetrical anymore. Completing the column-wise operation first prevents strided reading however this results in strided writing to the DRAM before row-wise traversal can start. This can be minimized by making efficient use of the local block RAM. Output columns are stored in the block RAM before being written to the DRAM in patches such that each row buffer access writes elements in several columns.

An alternative way to optimize the algorithm and reduce the resource consumption is to use equation (5.5) and take individual Fourier transforms of the row and column components. These can then be added with appropriate scaling terms. This optimization will reduce the FPGA resource consumption but will not improve the speed of the overall process, which is limited by the 2D FFT calculation.

5.7 Tile-Hopping Memory Mapping for 2D FFTs

In this section we propose a *tile-hopping* external memory access pattern for efficiently addressing external memory during intermediate storage between row-wise and column-wise 1D FFT operations to calculate a 2D FFT. As explained in section 5.4.2 and Fig. 5.3, column-wise reads from DRAM can be costly due to the over-head associated with activating and pre-charing. In the worst case scenario it can limit DRAM bandwidth up to 80% [143]. This is a problem with all such image processing operations where one stage of the processing has to be completed on all elements before the next stage can start. In the past there have been several implementations using local memory, however with growing demand for larger image sizes external memory has to be used. There have been several DRAM remapping attempts before such as [123, 130]. They propose a tile-based approach where a $n \times n$ image (input array) is divided into $\frac{n}{k} \times \frac{n}{k}$ tiles where k is the size of the DRAM row-buffer which allows for very high bandwidth DRAM access. Although, this method may be ideal to maximize the DRAM performance for 2D FFTs it incurs a high resource cost associated with local memory transposition and storing large chunks of data (entire row/column of tiles) in the local memory. Moreover, tiling in the image domain also requires remapping row-by-row operations. Another approach to reducing strided DRAM access has been presented in [144]. They present a 2D decomposition algorithm which decomposes the problem into smaller sub-block 2D FFTs which can be performed locally. This introduces extra row and column data exchanges and total number of operations are increased from $\mathcal{O}(n^2 \log n)$ to $\mathcal{O}(n^2(1 + \log n))$. Other implementations don't address the external memory issue in detail.

We propose *tile-hopping* address mapping which reduces the number of row activations required to access a single column. Unlike [123] our approach does not require significant local operations and storage and the row-by-row operations can proceed normally. The proposed memory mapping controller was designed on top of LabView FPGA's existing memory controller which efficiently controls interleaving and issues activation and pre-charge commands in parallel with data-transfer.

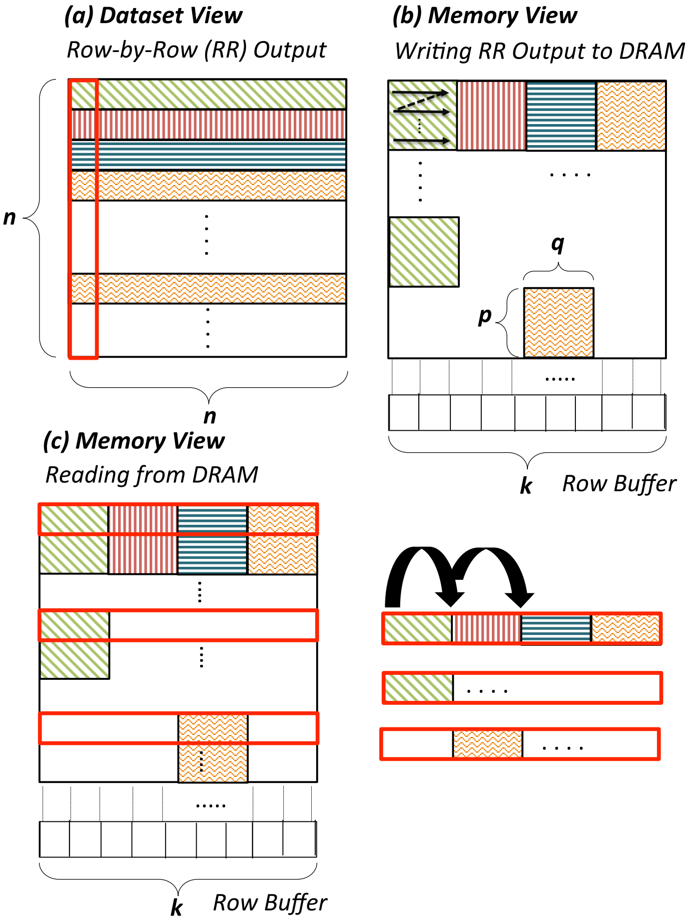


Figure 5.7: Image showing tile-hopping. a) Image-level view showing tiles. b) DRAM-level view showing tile placement while writing. c) DRAM-level view showing column reading from the tiles.

Instead of writing the results of the row-by-row 1D FFT in row-major order we remap the results in a blocked or tiled pattern as shown in Fig. 5.7. This means that when accessing an image column several elements of that column can be retrieved from a single DRAM row access. For a $n \times n$ image each row of size n can be divided into h tiles (*i.e.,* $n = h.N(t)$ where, $N(t)$ is the number of elements in each tile). These tiles can be remapped onto the DRAM "floor" as shown in Fig. 5.7. If the size of the row is small enough it may be possible to convert it into a single tile (*i.e.,* $h = 1$). However, this is unlikely for realistic image sizes. For a tile of size $p \times q$, a single row of the image is written into the DRAM by transitioning through $p.h$ rows. If k is the size of the row-buffer there are $\frac{k}{q}$ distinct tiles represented in each DRAM row and it contains same number of elements from a single image column. Given regular row-major storage when accessing column-wise elements one would have to transition through n DRAM rows to read a single image column. However, with this approach, when accessing an image column, $\frac{k}{q}$ elements of that column could be read from a single DRAM row which has been referred into the row buffer. Although, the cost of writing an image row is higher when compared to a standard row-major DRAM writing pattern, (*i.e.,* referring $p.h$ rather than n rows) the number of DRAM row referrals during column-wise read are reduced to $\frac{n.q}{k}$ which is $n.(1 - \frac{q}{k})$ less row referrals for a single column read.

We refer to this method as $tile - hopping$ because it entails mapping data onto several DRAM tiles and then hopping between the tiles such that several elements of the image column exist in a DRAM row which has been referred into the row bank. Although this mapping scheme has been developed for column-wise access required during 2D FFT calculation the scheme is general and can be adapted to other applications. Performance evaluation of this method has been presented in the experiments section.

5.8 Experimental Results and Analysis

5.8.1 Hardware Configuration and Target Selection

Since 2D DFTs are usually used for simplifying convolution operations in complex image processing and machine vision systems, we needed to prototype our design on a system that is expandable for next levels of processing. As mentioned earlier for rapid-prototyping of our proposed OPSD algorithm and tile-hopping memory mapping scheme we used a PXIe-based reconfigurable system. PXIe is an industrial extension of a PCI system with an enhanced bus structure that gives each connected device dedicated access to the bus with a maximum throughput of 24GB/s. This allows a high-speed dedicated link between a host PC and several FPGAs. The LabView FPGA graphical design environment is efficient for rapid-prototyping of complicated signal and image processing systems. It allows us to effectively integrate external HDL code and LabView graphical design on a single platform. Moreover, it allows a combination of high-level synthesis (HLS) and custom logic. Since current HLS tools have limitations when it comes to complex image and signal processing tasks, LabView FPGA tries to bridge these gaps by streamlining the design process.

We used FlexRIO (Flexible Reconfigurable I/O) FPGA boards plugged into a PXIe

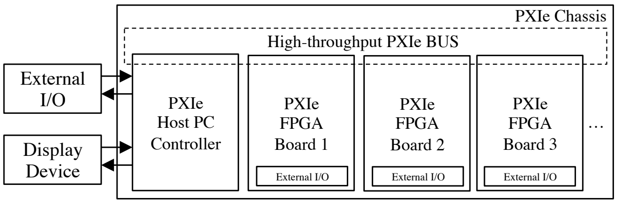


Figure 5.8: Block diagram of a PXIe based multi-FPGA system with a host PC controller connected through a high-speed bus on a PXIe chassis.

chassis. PXIe FlexRIO FPGA boards are adaptable and can be used to achieve high-throughput, because they allow direct data transfer between multiple FPGAs at rates as high as 8GB/s. This can significantly simplify multi-FPGA systems, that usually communicate via a host PC. This feature allows expansion of our system to further processing stages, making it flexible for a variety of applications. Fig. 5.8 shows a basic overview of a PXIe-based, multi-FPGA system with a host PC controller connected through a high-speed bus on a PXIe chassis.

Bandwidth Limitations

We used two NI PXIe-7976R FlexRIO board which have a Kintex 7 FPGA and 2GB external DRAM with theoretical data bandwidth up to 10GB/s. This FPGA board was plugged into a PXIe-1085 chassis along with a PXIe-8880 Intel Xeon PC controller. PXIe-1085 can hold up to 16 FPGAs and has 8 GB/s per-slot dedicated bandwidth and an overall system bandwidth of 24 GB/s.

5.8.2 Experimental Setup

As per Algorithms 5 and 6, discussed in previous sections, implementation involves four stages, A) Calculating the 2D FFT of a image frame. B) Calculating the Boundary Image. C) Calculating the 2D FFT of the Boundary image. D) Calculating the smooth component. E) Subtracting the smooth component from the 2D FFT of the original image to achieve the periodic component. The bottleneck consistently occurs in A and the serial part of the algorithm (B → C → D). The limitation due to A is reduced removing the so-called 'memory wall' by using our proposed *tile-hopping*-based memory mapping. The limitations due to the serial part of the algorithm are reduced by using OPSD rather than PSD.

The data flow presented in Fig. 5.5 was followed. Data-flow is clearly shown in a graphical programming environment making it easier to visualize how a design efficiently fits on an FPGA. Highly efficient implementations of 1D FFT were used from LabView FPGA for parallel row-by-row operations and by integrating Xilinx LogiCore for column-by-column operations. Each stage of the design was dynamically tested and benchmarked. The image was streamed from the host PC using a Direct Memory Access (DMA) FIFO. Eight rows of 1D FFTs are performed in parallel and stored in the DRAM via local memory in a tiled pattern as explained in the previous section. This follows reading several rows to extract a single column which is Fourier

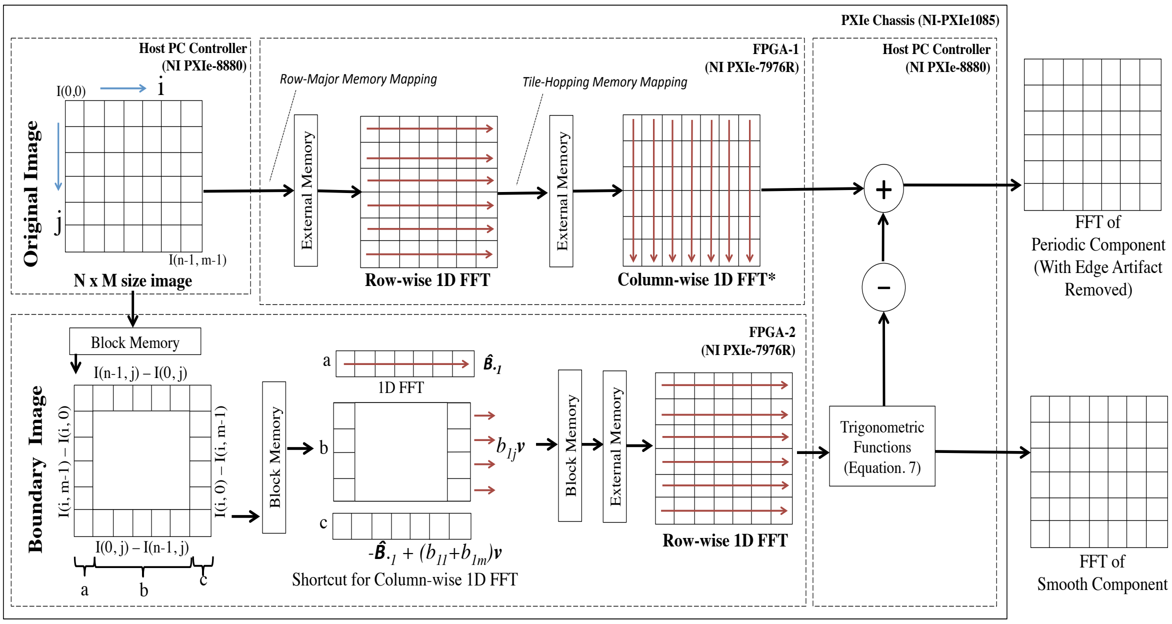


Figure 5.9: Functional block diagram of PXIe based 2D FFT implementation with simultaneous edge artifact removal using optimized periodic plus smooth decomposition. The OPSD algorithm is split among two NI-7976R (Kintex-7) FPGA boards with 2GB external memory and a host PC connected over a high-bandwidth bus. The image is streamed from the PC controller to FPGA 1 and FPGA 2. FPGA 1 calculates the row-by-row 1D FFT followed by column-by-column 1D FFT with intermediate tile-hopping memory mapping and sends the result back to the host PC. FPGA 2 receives the image, calculates the boundary image and proceeds to calculate the 1D FFT column-by-column FFT using the shortcut presented in Equation. 16 followed by row-by-row 1D FFTs and the result is sent back to the host PC.

transformed using Xilinx LogiCore and is sent back to the host PC. If the image is being streamed directly from an imaging device which scans and provides random or a non linear sequence of rows it is necessary to store a frame of the image in a buffer. This can also be accomplished by streaming the image flow from the host PC or using a smart camera which can delay image delivery by a single frame.

Data between the camera and the FPGA was transferred via the PC controller using Direct Memory Access (DMA) FIFO. However, it is also possible to directly transfer data to the FPGA by using a camera-link front-end. Local memory shown in Fig. 5.10 is used to buffer data between external memory and 1D FFT cores. This local memory is divided into read and write components and is implemented using FPGA slices. Block RAM (BRAM) is used for temporary storage of vectors required for calculating the 2D FFT of the boundary image (in the case of FPGA 2). The Control Unit (CU) organizes scheduling of transferring data between local and external memory. CU is based on LabView's existing memory controller and our memory mapping scheme presented in previous sections.

Step B was accomplished using standard LabView FPGA HLS tools using the graphical programming environment. In step C the 2D FFT of the boundary im-

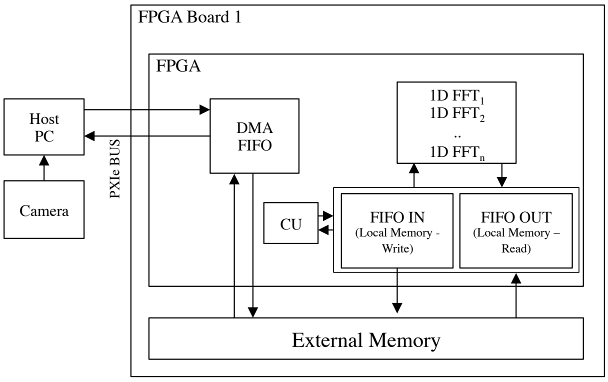


Figure 5.10: Block diagram of 2D FFT showing data transfer between external memory and local memory scheduled via a Control Unit (CU).

age needs to be calculated by row and column decomposition. However, as shown mathematically in the previous section, the initial row-wise FFTs can be calculated by computing the 1D FFT of the first (boundary) vector and the FFTs of remaining vectors can be computed by appropriate scaling of this vector. We need the boundary column vector for 1D FFT calculation of the first and last columns. We also need the boundary row vector for appropriate scaling of $\hat{\boldsymbol{v}}$ for the 1D FFT of every column between the first and last columns. Row and column vectors of the boundary image are stored in block RAM (BRAM). Fig. 5.9 shows a functional block diagram of the overall 2D FFT with optimized PSD process. Step D and E are performed on the host PC to minimize memory clashes and to access the periodic and smooth components of each frame as they become available.

5.8.3 Performance Evaluation

The overall performance of the system was evaluated using the setup presented in Fig. 5.9. The data was streamed from the host PC; in certain cases high frame rate videos as well as direct camera input was streamed from the host. All results presented are for 16-bit fixed point precision. Fig. 5.11a presents the effectiveness of our proposed tile-hopping memory mapping scheme. It clearly shows the effectiveness of our proposed memory mapping since it's closer to the theoretical peak performance. Fig. 5.11b presents the overall results comparing PSD and OPSD and demonstrating the effectiveness of our proposed optimization. PSD was also implemented on the same platform but the optimization presented in section IV was not used this rendered the serial portion of the algorithm to be the bottleneck which reduced overall performance. Table II shows a comparison of our implementation in contrast to recent 2D FFT FPGA implementations and shows that we achieve a better performance even with simultaneous edge artifact removal. Although, our implementation is tested with 16-bit fixed point precision which limits the accuracy of the transform, the precision may be sufficient for a variety of speed critical applications where alternative edge artifact removal methods (e.g. filtering etc.) may decrease overall system performance.

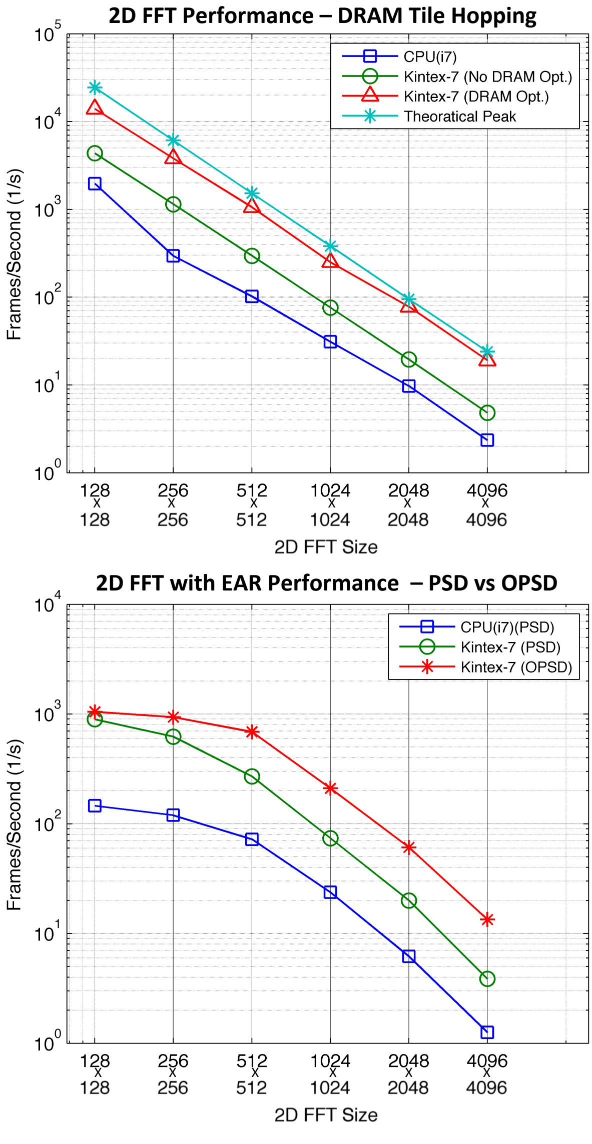


Figure 5.11: Performance evaluation in terms of frames per second for a) 2D FFTs with Tile-Hopping Memory Pattern. b) 2D FFTs with Edge Artifact Removal (EAR) using OPSD. The performance evaluation shows the significance of the two optimizations proposed. Both axes are on a log scale.

Table 5.2: Comparison of OPSD[1] 2D FFT with regular RCD-based implementations

Platform	SEAR[2]	Precision	Runtime	
	Yes/No	bits	512(ms)	1024(ms)
Kintex 7, 28nm (ours)	Yes	16 (fixed)	1.5	4.8
Kintex 7, 28nm (ours)	No	16 (fixed)	0.94	4.1
Stratix IV [123]	No	64 (double)	-	6.1
Virtex-5-BEE3, 65nm[131]	No	32 (single)	24.9	102.6
Virtex-E, 180nm [138]	No	16 (fixed)	28.6	76.9
ASIC, 180nm	No	32 (single)	21.0	-

[1] Optimized Periodic + Smooth Decomposition (PCD)

[2] Simultaneous Edge Artifact Removal

5.8.4 Energy Evaluation

The over-all energy consumption of the custom computing system depends on 1) power performance of the system components 2) throughput disparity. Throughput disparity results in idle time for at least one of the components and lowers the overall system throughput. A throughput-optimized system minimizes instances where certain components of the architecture are idle. The proposed optimizations in Section IV and Section V clearly reduce throughput disparity and minimize the ‘idle-time’ of the system. Thus, besides causing delays due to significant overhead standard column-wise DRAM access also contributes to the overall energy consumption. This is not only due to high count of DRAM row charges but is also because of energy consumed by the FPGA in idle state. Ideally, maximizing the DRAM bandwidth limits the amount of energy consumption. Proposed *’tile-hopping’* memory mapping scheme improves the DRAM bandwidth as seen in Fig. 10 and hence reduces the over-all energy consumption. The same is true for the proposed OPSD method where reduced DRAM access and 1D FFT invocations lead to reduced energy consumption. In this section we analyze the amount to improvement in energy consumption based on the proposed optimizations.

We estimate the DRAM power consumption for both the baseline (standard, strided memory access) as well as for the optimized (*’tile-hopping’* memory access) using MICRON DRAM power calculator. The energy is calculated in nJ for each read, *i.e.*, energy per read. This is accomplished by calculating the run-time for a specific 2D FFT and estimating the amount of energy consumed by the DRAM power calculator. Table 5.4 depicts the DRAM energy consumption for 2D FFTs before and after the proposed *’tile-hopping’* optimization for column-wise DRAM access. As mentioned earlier, row-wise DRAM access is fast and row-buffer size data can be accessed by a single row-activation. According to Table 5.4 the energy required for DRAM access is reduced by 42.7%, 48.8% and 52.9% for 1024×1024, 2048×2048 and 4096×4096 size 2D FFTs respectively.

The metric used to compare overall energy optimization achieved for 2D FFTs with EAR is energy per point, *i.e.*, the amount of average energy required to compute

Table 5.3: DRAM Energy Consumption Baseline vs Tile-Hopping

	1024×1024	2048×2048	4096×4096
	nJoule	*nJoule*	*nJoule*
EPR[*] CW[°] Read (Baseline)	4.46	5.77	7.12
EPR CW Read *Tile-Hopping* **(Proposed)**	2.54	2.95	3.36
Reduction (%)	**42.7%**	**48.8%**	**52.9%**

[*]Energy per read (EPR).
[°]Column-wise memory access (CW).

Table 5.4: 2D FFT + EAR Energy Consumption Baseline vs Optimized (OPSD+Tile Hopping)

	1024×1024	2048×2048	4096×4096
	nJoule	*nJoule*	*nJoule*
EPP[†] 2D FFT+EAR Baseline	36.92	41.25	48.35
2D FFT+EAR (Opt.) *Tile-Hopping+OPSD* **(Proposed)**	15.88	17.06	18.11
Improvement	2.3×	2.4×	2.8×

[†]Energy per point (EPP).

the 2D FFT of a single point in an image with simultaneous edge artifact removal. This was achieved by calculating the energy consumed by Xilinx LogiCORE IP for 1D FFTs, the DRAM and the edge artifact removal part separately. The estimated energy calculated does not include energy consumed by the PXIe chassis and the host PC. Essentially, the FPGA-based architecture presented here could be used without the host controller. The energy consumption incorporates dynamic as well as static power. The overall energy consumption per point is reduced by 56.9%, 58.6% and 62% for calculating 1024×1024, 2048×2048 and 4096×4096 size 2D FFTs with EAR respectively.

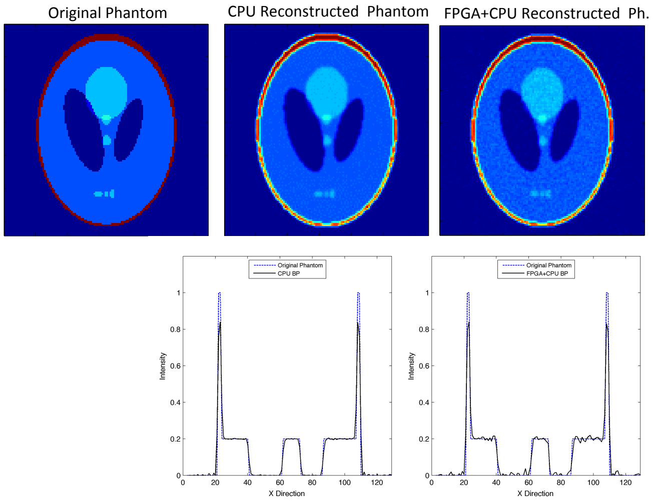


Figure 5.12: Figure showing a thin slice of filtered back-projection results by reconstructing a 128 × 128 × 128 Shepp-Logan phantom. The 3D density was reconstructed from 180 equally spaced simulated projections using standard linearly interpolated FBP and using a ram-lak filter. It can be seen that the results from the FPGA+CPU solution have some errors this is due to the fact that the 2D FFT of each projection is less accurate. The results are good enough to be used as a prior for further optimization based refinement methods.

5.9 Application: Filtered Back-projection for Tomography

In order to further demonstrate the effectiveness of our implementation we use the created 2D FFT module as an accelerator for reducing the run-time for filtered back-projection (FBP). In depth details regarding the basic FBP algorithm have been discussed in Chapter 1, but can be found in [10, 21]. The method can be used to reconstruct primitive 3D tomograms from 2D data, which can then be used as a prior for more complex regularization-based methods such as [15, 67, 145]. The algorithmic flow is based on the Fourier slice theorem, *i.e.,* 2D Fourier transforms of projections are an angular component of the 3D Fourier transform of the 3D reconstructed volume. Our 2D FFT accelerator was used to calculate the 2D FFTs of the projections as well as for initial stages of the 3D FFT which was then completed on the host PC. Similar to the 2D FFT the 3D FFT is separable and can be divided into 2D FFTs and 1D FFTs. The results have been shown in Table 5.5, it can be seen that the improvement for smaller size densities is not significant because their FFTs are quite fast on general purpose CPUs. However, for larger densities the FFT accelerator can give a significant improvement. If the remaining components are also implemented on an FPGA, significant speed increase can be achieved. Results of a thin slice from a 3D simulated shepp-logan [26] phantom have been shown in Fig. 5.12. It can be seen that the result from the hardware accelerated FBP are of slightly lower quality. This is due to the fact that our 2D FFT implementation is less accurate (16 bit, fixed-point) as compared to the CPU-based implementation (FFTW, double-precision floating). The accelerated FBP was also tested with real Electron Tomography (ET) data.

Table 5.5: Comparing Filtered Back-projection Runtime

3D Density	CPU (i7)	FPGA + Host PC (i7)
	Sec	Sec
128 × 128 × 128	21.3 sec	19.5 sec
256 × 256 × 256	47.5 sec	42.4 sec
512 × 512 × 512	94.8 sec	81.3 sec
1024 × 1024 × 1024	322.3 sec	275.3 sec
2048 × 2048 × 2048	1687.7 sec	1364.4 sec
4096 × 4096 × 4096	16463.1 sec	12599.4 sec

5.10 Conclusion

2D FFTs often become a major bottleneck for high-performance imaging and vision systems. The inherent computational complexity of the 2D FFT kernel is further enhanced if effective removal (using PSD) of spurious artifacts introduced by the non-periodic nature of real-life images is taken into account. We developed and implemented an FPGA-based design for calculating high-throughput 2D DFTs with simultaneous edge artifact removal. Our approach is based on a PSD algorithm that splits the frequency domain of a 2D image into a smooth component, which contains the high-frequency, cross-shaped artifacts and can be subtracted from the 2D DFT of the original image to obtain a periodic component that is artifact free. Since this approach calculates two 2D DFTs simultaneously, external memory addressing and repeated 1D FFT invocations become problematic. To solve this problem we optimized the original PSD algorithm to reduce the number of DFT samples to be computed and DRAM access by 24%. Moreover, to reduce strided access from the DRAM during column-wise reads we presented and analyzed *'tile-hopping'*, a memory mapping scheme which reduces the number of DRAM row activation when reading a single column of data. This memory mapping scheme is general and may be used for a variety of other applications.

Our methods were tested using extensive synthesis and benchmarking using a Xilinx Kintex 7 FPGA communicating with a host PC on a high-speed PXIe bus. Our system is expandable to support several FPGAs and can be adapted to various large-scale computer vision and biomedical applications. Despite decomposing the image into periodic and smooth frequency components our design requires less run-time, compared to traditional FPGA-based 2D DFT implementations and can be used for a variety of highly demanding applications. One such application, filtered back-projection was accelerated using the proposed implementation to achieve better results specifically for larger size raw tomographic data.

Chapter 6

Conclusions and Future Directions

6.1 Summary and Conclusions

This thesis focused on the problem of refining Cryo-ET reconstructions from an image reconstruction point of view. Cryo-ET despite being a powerful structure determination method becomes an *ill-posed* inverse reconstruction problem, made tedious by various aspects of data collection constraints. The thesis also focuses on reconfigurable computing as a medium for accelerating Cryo-ET reconstructions in specific and tomographic reconstructions in general.

A review of Cryo-ET methodology and fundamentals of tomography was presented in Chapter 1. Cryo-ET is a macromolecular structure determination method which is better than other methods in the sense that it allows samples to be imaged in their native sates. Due to rapid freezing, and less prominent requirement of averaging Cryo-ET can preserve information pertaining to the flexibility of the molecule being imaged giving the possibility to study molecular dynamics. The drawbacks of Cryo-ET involve lower resolution which can be remedied to some extent by sub-tomogram averaging. Cryo-ET data collection suffers from a wide variety of problems which stem from the missing wedge, the notorious dose problem, and detector imperfections etc. These issues render the inverse reconstruction problem ill-posed, *i.e.*, it lacks a unique solution and is unstable. Solving such an ill-posed inverse problem can be a challenge and the problem ties into the field of image reconstruction more specifically iterative image reconstruction which is an active research area. One of the major objectives of this thesis was to develop, analyze and study algorithms which could overcome some of the issues associated with the Cryo-ET reconstruction problem.

In regards to this Chapter 2 presented a sinogram denoising a pre-processing based method. This method makes use of the fact that the sinogram has structured data due to high redundancy in the projections since they are collected from the same sample. Employing non-local graph-based methods it is possible to make use of redundancy in images since they establish connections between regions of the image which are spatially far from each other. Since the sinogram has proximal but not necessarily adjacent pixels which are similar we could reduce the nearest neighbor search parameter. It was shown via extensive experimentation that it is possible to denoise the sinogram and achieve better reconstructions using standard reconstruction methods such as FBP, ART and SIRT etc.

This non-local graph-based denoising work inspired Chapter 3 where graph-based non-local denoising was incorporated in the reconstruction method to achieve better reconstructions. Chapter 3 dives into CS-based SEIR methods which harness the concepts of compressing to acquired data, *i.e.*, these methods treat the acquired data as if it was a compressed form of data in some sparse bases. Usually CS alone is not enough to cater for the sparsity required for real problems. Therefore the proposed method (AGTV) promotes sparsity in the wavelet and graph-gradient domains. AGTV was extensively tested in different conditions and it was shown that parts of the missing wedge could be recovered by employing this method. That said, the computational constraints associated with the method are significant. However, the method is of immense interest for the image reconstruction community and can be seen as a generalization of the state-of-the-art CSTV and NLTV/CSGT methods. The method acts as a major proof-of-concept and an inspiration for future studies along these lines.

The computational complexity associated with non-local methods is significant hence Chapter 4 focuses on a more efficient and intuitive method to denoise tomograms. Extended reconstructions work on the concept that reconstructing a larger 'artificial' region outside the ROI during regularization-based image reconstruction gives 'lee-way' for the noise to spread out throughout the reconstruction space while the much more confined signal stays within the ROI. The signal has a clear representation in every projection, while this is not the case for noise and inconsistency errors which can spread to this extra region. This concept was tested with extensive simulation experiments as well as real Cryo-ET 3D data. Multimedia files related to this method have been included with this thesis.

Finally, Chapter 5 focuses on reconfigurable computing and FPGA-based acceleration of 2D FFTs with edge artifact removal. This is a fundamental component of Fourier slice theorem-based image reconstruction methods. From a tomographic point of view this is interesting because FBP is used as a prior for most iterative image reconstruction methods and is subject to the ROI problem in tomography. The ROI problem occurs when an extracted region from a tilt series of projections is used rather than the entire projections. In this case each extraction has information regarding the region which is not a part of the ROI. This can partially be reduced by reconstructing the entire size of the projections. However, this is computationally tedious even for a primitive method like FBP. This serves the major motivation behind taking the time and spending the effort to implement high performance 2D FFTs on FPGAs. FBP has traditionally been implemented in the real-space but for large size implementations real-space will require a lot of data shuffling something not trivial to accomplish on semi-high-level FPGA implementation environments. Moreover, 2D FFT FPGA implementation on an expandable system is only the first step towards a long term objective of implementing full iterative methods in a reconfigurable computing setting. Also, the 2D FFT is the first step towards implementing a NUFFT, a method commonly used in tomography since it can avoid the requirement of Fourier interpolation. 2D FFTs have a major issue related to edge artifacts which stems from the fact that the edges of an image are not periodic, the 2D FFT implementation presented in Chapter 4 incorporates an optimized edge artifact removal scheme with minimal loss of frequency information (which is required for tomographic applications). The architecture presented in Chapter 4 is general and can be adapted to other types of hardware.

6.2 Original Contributions

The following original contributions have been made in this thesis:

From an Image Reconstruction Perspective

- The proposition of graph-based non-local sinogram denoising, its analysis and effectiveness for denoising tomographic reconstruction (Chapter 2).
- The proposition of adaptive graph-based total variation as a compressed sensing-type method which can simultaneously reconstruct and denoise data. The ability of the method to reconstruct from missing and noisy data specifically the recovery of data from the missing wedge. A unique aspect of the method is that it can be seen as a generalization of CS and TV-type methods, a fact of significance and interest for the image reconstruction community (Chapter 3).
- Analysis of extended field-based reconstructions as a method that can be applied to real data and is computationally efficient. Although, the idea of this method was first presented in [105] in 1974 at that time the method was only applied to 2D data[1]. The major contribution in this area was to enable the use of extended field with a wide variety of methods and to study its properties pertaining to the extension size and its ability to reduce the regularization parameter (Chapter 5).

From a Reconfigurable Computing Perspective

- Development of an FPGA-based 2D FFT architecture with simultaneous edge artifact removal for high-performance applications. This was achieved by making the following contributions.
- Optimizing the existing periodic plus smooth-based edge artifact removal scheme to reduce the access of DRAM and 1D FFT invocations.
- Designing a custom memory controller to overcome the DRAM column-wise access issue which results in reduced memory access speeds while accessing data stored column-wise in the memory.

6.3 Future Directions

This thesis is a good proof-of-concept and can be seen as an initial study specifically in the context of non-local graph-based methods for tomographic refinement. However, there are some questions that have not been addressed by this thesis and many more that come into mind as result of the studies presented here. The collective aim of the thesis was the improvement and refinement of tomographic reconstructions. The following future studies can follow as a result of the developments presented here:

[1] A fact that I find intriguing is that the 1974 method was computationally complex at the time for them to test with 3D real data something which is tedious for us today in the case of graph-based non-local methods.

- Sinogram denoising presented in Chapter 2 is only an initial proof of concept, the study can be extended for 3D data by using the recently developed framework for tensors on graphs [146]. Although, this might not be very simple since the optimization problem will become increasingly complex and a forward-backward primal dual-based solver used in the current study may not be applicable.

- AGTV presented in Chapter 3 has several constraints including computational constraints, hyper-parameter tuning and the curse of adaptivity. Several subsequent studies could be designed to reduce the computational complexity of graph construction and solving the optimization problem. In this regard one approach could be to move to a Chambolle-Pock [147] Optimization scheme rather than a proximal forward-backward primal dual-based approach. However, this may not be very simple since extensive mathematical analysis of the problem may be required.

- A more detailed mathematical analysis of Extended Field-based tomographic reconstructions may lead to interesting details of how the method can be incorporated within the optimization process of variational regularization methods so as to preserve the fit with the original data.

- The 2D FFT architecture presented in Chapter 5 is simply the first step towards implementing much more complex iterative and optimization-based methods for tomography in a reconfigurable computing-based setting. The setup has been created in a way that one can easily extend the PXIe system by using more FPGAs and communicating between them over a high-speed bus. That said, it is not trivial to accomplish large scale FPGA tasks even within the framework of semi-high-level design.

Appendix A

Details Regarding Multimedia Files

Multimedia File - I: RealData.mp4
Description: The video shows a $355 \times 355 \times 255$ COMET reconstruction of BENDZIL colloidal silica and several extended COMET reconstructions ranging from $355 \times 355 \times 305$ to $355 \times 355 \times 555$. Each successive reconstruction is an extract from an extended reconstruction larger than the ROI. It can clearly be seen that the larger the reconstruction, more of the noise is redistributed to the extended region rather than the ROI.
Size: 9.7MB
Player Information: The video was created using iMove from video files generated by screen video capture from 3D visualization software Brick of Bytes (BOB). The video uses codec H.264 and can be viewed on QuickTime Player on MAC or Windows Media Player 12 / VCL Player on Windows.

Multimedia File - II: ART-Binary.gif
36 projections were generated from 32×32 Binary Phantom and the resulting matrix of projections (sinogram) was corrupted with 20% added normalized noise. The noise corrupted sinogram was then extended by zero-padding corresponding to a 64×64 reconstruction space. The resulting sinogram was then reconstructed using ART (Kaczmarz Method). As evident from the multimedia file the resulting reconstruction has an extra extended region for the anomalous noise to spread into. Kaczmarz method was run for 100 iterations and achieved semi-convergence at the 15^{th} iteration and had an error of 7.51. The multimedia file has 15 frames each corresponding to an iteration of Kaczmarz.
Size: 747KB
Player Information: The file can be viewed in any web browser.

Multimedia File - III: ART-Shepp-Logan.gif
36 projections were generated from 64×64 Shepp-Logan Phantom and the resulting matrix of projections (sinogram) was corrupted with 20% added normalized noise. The noisy sinogram was then extended by zero-padding corresponding to a 128×128 reconstruction space. The resulting sinogram was then reconstructed using ART (Kaczmarz Method). As evident from the multimedia file the resulting reconstruction has an extra extended region for the anomalous noise to spread into. Kaczmarz method was run for

100 iterations and it achieved semi-convergence at the 5^{th} iteration and had an error of 9.78. The multimedia file has 5 frames each corresponding to an iteration of Kaczmarz.
Size: 290KB
Player Information: The file can be viewed in any web browser.

Multimedia File - IV: SIRT-Smooth.gif
36 projections were generated from 64×64 Smooth Phantom and the resulting matrix of projections (sinogram) was corrupted with 8% added normalized noise. This noisy sinogram was then extended by zero-padding corresponding to a 128×128 reconstruction space. The resulting sinogram was then reconstructed using SIRT (Cimmino). As evident from the multimedia file the resulting reconstruction has an extra extended region for the anomalous noise to spread into. Cimmino's method was run for 100 iteration and achieved it achieved semi-convergence at the 7^{th} iteration and had an error of 3.11. The multimedia file has 7 frames each corresponding to an iteration of Cimmino.
Size: 1.7MB
Player Information: The file can be viewed in any web browser.

Multimedia File - V: SIRT-Shepp-Logan.gif
36 projections were generated from 64×64 Binary Phantom and the resulting matrix of projections (sinogram) was corrupted with 10% added normalized noise. This noisy sinogram was then extended by zero-padding corresponding to a 128×128 reconstruction space. The resulting sinogram was then reconstructed using SIRT (Cimmino). As evident from the multimedia file the resulting reconstruction had an extra extended region for the anomalous noise to spread into. Cimmino's method was run for 100 iterations and it achieved semi-convergence at the 56^{th} iteration and has an error of 5.94. The multimedia file has 56 frames each corresponding to an iteration of Cimmino.
Size: 4.4MB
Player Information: The file can be viewed in any web browser.

Multimedia File - VI: 2D FFT - EAR - Demo.mp4
Video showing 320×240 2D FFT with simultaneous edge artifact removal of a fractal pattern.
Size: 3.6MB
Player Information: The video was created using iMove from video files generated by screen video capture from 3D visualization software Brick of Bytes (BOB). The video uses codec H.264 and can be viewed on QuickTime Player on MAC or Windows Media Player 12 / VCL Player on Windows.

Bibliography

[1] J. Elands and U. Skoglund, "Cryo electron microscopy and electron tomography will play a crucial role in the future of drug development," *Drug Discovery*, p. 81, 2005.

[2] H. R. Saibil, "Macromolecular structure determination by cryo-electron microscopy," *Acta Crystallographica Section D: Biological Crystallography*, vol. 56, no. 10, pp. 1215–1222, 2000.

[3] A. E. Todd, R. L. Marsden, J. M. Thornton, and C. A. Orengo, "Progress of Structural Genomics Initiatives: An Analysis of Solved Target Structures," *Journal of Molecular Biology*, vol. 348, no. 5, pp. 1235–1260, May 2005.

[4] J. Frank, *Electron tomography: methods for three-dimensional visualization of structures in the cell.* Springer Science & Business Media, 2008.

[5] S. Sandin, L.-G. Öfverstedt, A.-C. Wikström, Wrange, and U. Skoglund, "Structure and Flexibility of Individual Immunoglobulin G Molecules in Solution," *Structure*, vol. 12, no. 3, pp. 409–415, Mar. 2004.

[6] A. J. Koster, R. Grimm, D. Typke, R. Hegerl, A. Stoschek, J. Walz, and W. Baumeister, "Perspectives of Molecular and Cellular Electron Tomography," *Journal of Structural Biology*, vol. 120, no. 3, pp. 276–308, Dec. 1997.

[7] Y. Censor, "Finite series-expansion reconstruction methods," *Proceedings of the IEEE*, vol. 71, no. 3, pp. 409–419, 1983.

[8] U. Skoglund and B. Daneholt, "Electron microscope tomography," *Trends in Biochemical Sciences*, vol. 11, no. 12, pp. 499–503, 1986.

[9] A. Leis, M. Beck, M. Gruska, C. Best, R. Hegerl, and J. Leis, "Cryo-electron tomography of biological specimens," *IEEE Signal Processing Magazine*, vol. 23, no. 3, pp. 95–103, May 2006.

[10] R. A. Crowther, D. J. DeRosier, and A. Klug, "The Reconstruction of a Three-Dimensional Structure from Projections and its Application to Electron Microscopy," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 317, no. 1530, pp. 319–340, Jun. 1970.

[11] D. Fanelli and O. Öktem, "Electron tomography: a short overview with an emphasis on the absorption potential model for the forward problem," *Inverse Problems*, vol. 24, no. 1, p. 013001, Feb. 2008.

[12] O. Öktem, "Mathematics of Electron Tomography," *Handbook of Mathematical Methods in Imaging*, pp. 937–1031, 2015.

[13] L. A. Baker and J. L. Rubinstein, "Chapter Fifteen-Radiation Damage in Electron Cryomicroscopy," *Methods in enzymology*, vol. 481, pp. 371–388, 2010.

[14] R. Egerton, P. Li, and M. Malac, "Radiation damage in the TEM and SEM," *Micron*, vol. 35, no. 6, pp. 399–409, 2004.

[15] U. Skoglund, L.-G. Öfverstedt, R. M. Burnett, and G. Bricogne, "Maximum-Entropy Three-Dimensional Reconstruction with Deconvolution of the Contrast Transfer Function: A Test Application with Adenovirus," *Journal of Structural Biology*, vol. 117, no. 3, pp. 173–188, Nov. 1996.

[16] J.-J. Fernandez, "Computational methods for electron tomography," *Micron*, vol. 43, no. 10, pp. 1010–1030, Oct. 2012.

[17] Z. Saghi, G. Divitini, B. Winter, R. Leary, E. Spiecker, C. Ducati, and P. A. Midgley, "Compressed sensing electron tomography of needle-shaped biological specimens – Potential for improved reconstruction fidelity with reduced dose," *Ultramicroscopy*, vol. 160, pp. 230–238, Jan. 2016.

[18] P. C. Hansen, *Discrete inverse problems: insight and algorithms.* Siam, 2010, vol. 7.

[19] S. Webb, *From the watching of shadows: the origins of radiological tomography.* CRC Press, 1990.

[20] S. Helgason, "The Radon Transform on R n," in *Integral Geometry and Radon Transforms.* Springer, 2011, pp. 1–62.

[21] F. Natterer, *The mathematics of computerized tomography.* Siam, 1986, vol. 32.

[22] J. Frikel and E. T. Quinto, "Characterization and reduction of artifacts in limited angle tomography," *Inverse Problems*, vol. 29, no. 12, p. 125007, 2013.

[23] R. Narasimha, I. Aganj, A. E. Bennett, M. J. Borgnia, D. Zabransky, G. Sapiro, S. W. McLaughlin, J. L. Milne, and S. Subramaniam, "Evaluation of denoising algorithms for biological electron tomography," *Journal of structural biology*, vol. 164, no. 1, pp. 7–17, 2008.

[24] J. August Radon, "On determination of functions by their integral values along certain multiplicities," *Ber. der Sachische Akademie der Wissenschaften Leipzig,(Germany)*, vol. 69, pp. 262–277, 1917.

[25] J. Radon, "On the determination of functions from their integral values along certain manifolds," *IEEE transactions on medical imaging*, vol. 5, no. 4, pp. 170–176, 1986.

[26] L. A. Shepp and B. F. Logan, "The Fourier reconstruction of a head section," *IEEE Transactions on Nuclear Science*, vol. 21, no. 3, pp. 21–43, Jun. 1974.

[27] R. N. Bracewell, “Strip integration in radio astronomy,” *Australian Journal of Physics*, vol. 9, no. 2, pp. 198–217, 1956.

[28] A. C. Kak and M. Slaney, *Principles of computerized tomographic imaging.* IEEE press, 1988.

[29] Y. Censor, D. Gordon, and R. Gordon, “Component averaging: An efficient iterative parallel algorithm for large and sparse unstructured problems,” *Parallel Computing*, vol. 27, no. 6, pp. 777–808, May 2001.

[30] A. Gopinath, G. Xu, D. Ress, O. Oktem, S. Subramaniam, and C. Bajaj, “Shape-based regularization of electron tomographic reconstruction,” *IEEE transactions on medical imaging*, vol. 31, no. 12, pp. 2241–2252, 2012.

[31] P. C. Hansen and M. Saxild-Hansen, “AIR tools—a MATLAB package of algebraic iterative reconstruction methods,” *Journal of Computational and Applied Mathematics*, vol. 236, no. 8, pp. 2167–2178, 2012.

[32] G. Moore, “Cramming more components onto integrated circuits,” *Readings in computer architecture*, vol. 56, 2000.

[33] G. Estrin, B. Bussell, R. Turn, and J. Bibb, “Parallel processing in a restructurable computer system,” *IEEE Transactions on Electronic Computers*, no. 6, pp. 747–755, 1963.

[34] R. Hartenstein, “A decade of reconfigurable computing: a visionary retrospective,” in *Proceedings of the conference on Design, automation and test in Europe.* IEEE Press, 2001, pp. 642–649.

[35] K. Compton and S. Hauck, “Reconfigurable computing: a survey of systems and software,” *ACM Computing Surveys (csuR)*, vol. 34, no. 2, pp. 171–210, 2002.

[36] F. Mahmood, N. Shahid, P. Vandergheynst, and U. Skoglund, “Graph Based Sinogram Denoising for Tomographic Reconstructions,” *IEEE International Conference on Engineering in Medical and Biology 2016*, 2016.

[37] J. Hsieh, “Computed tomography: principles, design, artifacts, and recent advances.” SPIE Bellingham, WA, 2009.

[38] D. Karimi, P. Deman, R. Ward, and N. Ford, “A sinogram denoising algorithm for low-dose computed tomography,” *BMC Medical Imaging*, vol. 16, no. 1, Dec. 2016.

[39] J. A. Fessler, “Statistical image reconstruction methods for transmission tomography,” *Handbook of medical imaging*, vol. 2, pp. 1–70, 2000.

[40] J. Wang, H. Lu, T. Li, and Z. Liang, “Sinogram noise reduction for low-dose CT by statistics-based nonlinear filters,” Apr. 2005, pp. 2058–2066.

[41] A. Björck and T. Elfving, "Accelerated projection methods for computing pseudoinverse solutions of systems of linear equations," *BIT Numerical Mathematics*, vol. 19, no. 2, pp. 145–163, 1979.

[42] J. Hsieh, "Adaptive streak artifact reduction in computed tomography resulting from excessive x-ray photon noise," *Medical Physics*, vol. 25, no. 11, pp. 2139–2147, Nov. 1998.

[43] M. Beister, D. Kolditz, and W. A. Kalender, "Iterative reconstruction methods in X-ray CT," *Physica Medica*, vol. 28, no. 2, pp. 94–108, Apr. 2012.

[44] A. Buades, B. Coll, and J. M. Morel, "A Review of Image Denoising Algorithms, with a New One," *Multiscale Modeling & Simulation*, vol. 4, no. 2, pp. 490–530, Jan. 2005.

[45] D. Karimi and R. K. Ward, "Sinogram denoising via simultaneous sparse representation in learned dictionaries," *Physics in medicine and biology*, vol. 61, no. 9, p. 3536, 2016.

[46] P. J. La Rivière, "Penalized-likelihood sinogram smoothing for low-dose CT," *Medical Physics*, vol. 32, no. 6, p. 1676, 2005.

[47] P. La Riviere, Junguo Bian, and P. Vargas, "Penalized-likelihood sinogram restoration for computed tomography," *IEEE Transactions on Medical Imaging*, vol. 25, no. 8, pp. 1022–1036, Aug. 2006.

[48] J. Wang, H. Lu, Z. Liang, D. Eremina, G. Zhang, S. Wang, J. Chen, and J. Manzione, "An experimental study on the noise properties of x-ray CT sinogram data in Radon space," *Physics in Medicine and Biology*, vol. 53, no. 12, pp. 3327–3341, Jun. 2008.

[49] M. Kachelrieß, O. Watzke, and W. A. Kalender, "Generalized multi-dimensional adaptive filtering for conventional and spiral single-slice, multi-slice, and cone-beam CT," *Medical Physics*, vol. 28, no. 4, pp. 475–490, Apr. 2001.

[50] J. L. Prince and J. M. Links, *Medical imaging signals and systems.* Pearson Prentice Hall Upper Saddle River, NJ, 2006.

[51] A. Buades, B. Coll, and J.-M. Morel, "Non-local means denoising," *IEEE Signal Processing Magazine*, vol. 1, 2011.

[52] G. Peyré, S. Bougleux, and L. Cohen, "Non-local regularization of inverse problems," in *European Conference on Computer Vision.* Springer, 2008, pp. 57–68.

[53] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The Emerging Field of Signal Processing on Graphs: Extending High-Dimensional Data Analysis to Networks and Other Irregular Domains," *arXiv preprint arXiv:1211.0053*, 2012.

[54] D. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, “Signal Processing on Graphs,” *Signal Processing*, no. 2/35, 2013.

[55] A. Buades, B. Coll, and J.-M. Morel, “A Non-Local Algorithm for Image Denoising,” vol. 2. IEEE, 2005, pp. 60–65.

[56] L. I. Rudin, S. Osher, and E. Fatemi, “Nonlinear total variation based noise removal algorithms,” *Physica D: Nonlinear Phenomena*, vol. 60, no. 1, pp. 259–268, 1992.

[57] P. L. Combettes and J.-C. Pesquet, “Proximal splitting methods in signal processing,” in *Fixed-point algorithms for inverse problems in science and engineering.* Springer, 2011, pp. 185–212.

[58] N. Perraudin, D. Shuman, G. Puy, and P. Vandergheynst, “UNLocBoX A matlab convex optimization toolbox using proximal splitting methods,” *arXiv preprint arXiv:1402.0779*, 2014.

[59] R. A. Brooks and G. Di Chiro, “Theory of Image Reconstruction in Computed Tomography [1],” *Radiology*, vol. 117, no. 3, pp. 561–572, Dec. 1975.

[60] R. Gordon, R. Bender, and G. T. Herman, “Algebraic Reconstruction Techniques (ART) for three-dimensional electron microscopy and X-ray photography,” *Journal of Theoretical Biology*, vol. 29, no. 3, pp. 471–481, Dec. 1970.

[61] P. Gilbert, “Iterative methods for the three-dimensional reconstruction of an object from projections,” *Journal of Theoretical Biology*, vol. 36, no. 1, pp. 105–117, Jul. 1972.

[62] G. Cimmino and C. N. delle Ricerche, *Calcolo approssimato per le soluzioni dei sistemi di equazioni lineari.* Istituto per le applicazioni del calcolo, 1938.

[63] A. Wirgin, “The inverse crime,” *arXiv preprint math-ph/0401050*, 2004.

[64] J. Kaipio and E. Somersalo, “Statistical inverse problems: discretization, model reduction and inverse crimes,” *Journal of computational and applied mathematics*, vol. 198, no. 2, pp. 493–504, 2007.

[65] H. Rullgård, L.-G. Öfverstedt, S. Masich, B. Daneholt, and O. Öktem, “Simulation of transmission electron microscope images of biological specimens,” *Journal of microscopy*, vol. 243, no. 3, pp. 234–256, 2011.

[66] R. Leary, Z. Saghi, P. A. Midgley, and D. J. Holland, “Compressed sensing electron tomography,” *Ultramicroscopy*, vol. 131, pp. 70–91, Aug. 2013.

[67] F. Mahmood, N. Shahid, U. Skoglund, and P. Vandergheynst, “Adaptive Graph-based Total Variation for Tomographic Reconstructions,” *arXiv preprint arXiv:1610.00893*, 2016.

[68] A. Berrington de González, “Projected Cancer Risks From Computed Tomographic Scans Performed in the United States in 2007,” *Archives of Internal Medicine*, vol. 169, no. 22, p. 2071, Dec. 2009.

[69] D. J. Brenner and E. J. Hall, “Computed Tomography—An Increasing Source of Radiation Exposure,” *N Engl J Med*, vol. 357, pp. 2277–84, 2007.

[70] M. S. Pearce, J. A. Salotti, M. P. Little, K. McHugh, C. Lee, K. P. Kim, N. L. Howe, C. M. Ronckers, P. Rajaraman, A. W. Craft, and others, “Radiation exposure from CT scans in childhood and subsequent risk of leukaemia and brain tumours: a retrospective cohort study,” *The Lancet*, vol. 380, no. 9840, pp. 499–505, 2012.

[71] E. T. Quinto, U. Skoglund, and O. Öktem, “Electron lambda-tomography,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 51, pp. 21 842–21 847, 2009.

[72] Y. Censor, “Row-Action Methods for Huge and Sparse Systems and Their Applications,” *SIAM Review*, vol. 23, no. 4, pp. 444–466, Oct. 1981.

[73] J. Qi and R. M. Leahy, “Iterative reconstruction techniques in emission computed tomography,” *Physics in medicine and biology*, vol. 51, no. 15, p. R541, 2006.

[74] H. Rullgård, O. Öktem, and U. Skoglund, “A componentwise iterated relative entropy regularization method with updated prior and regularization parameter,” *Inverse Problems*, vol. 23, no. 5, pp. 2121–2139, Oct. 2007.

[75] H. M. Hudson and R. S. Larkin, “Accelerated image reconstruction using ordered subsets of projection data,” *IEEE transactions on medical imaging*, vol. 13, no. 4, pp. 601–609, 1994.

[76] H. Erdogan and J. A. Fessler, “Ordered subsets algorithms for transmission tomography,” *Physics in medicine and biology*, vol. 44, no. 11, p. 2835, 1999.

[77] H. Nien and J. A. Fessler, “Fast X-ray CT image reconstruction using a linearized augmented Lagrangian method with ordered subsets,” *IEEE transactions on medical imaging*, vol. 34, no. 2, pp. 388–399, 2015.

[78] D. L. Donoho, “Compressed sensing,” *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[79] G.-H. Chen, J. Tang, and S. Leng, “Prior image constrained compressed sensing (PICCS): A method to accurately reconstruct dynamic CT images from highly undersampled projection data sets,” *Medical Physics*, vol. 35, no. 2, p. 660, 2008.

[80] C. G. Graff and E. Y. Sidky, “Compressive sensing in medical imaging,” *Applied Optics*, vol. 54, no. 8, p. C23, Mar. 2015.

[81] J. Song, Q. H. Liu, G. A. Johnson, and C. T. Badea, "Sparseness prior based iterative image reconstruction for retrospectively gated cardiac micro-CT," *Medical physics*, vol. 34, no. 11, pp. 4476–4483, 2007.

[82] L. Ritschl, F. Bergner, C. Fleischmann, and M. Kachelries, "Improved total variation-based CT image reconstruction applied to clinical data," *Physics in medicine and biology*, vol. 56, no. 6, p. 1545, 2011.

[83] J. Tang, B. E. Nett, and G.-H. Chen, "Performance comparison between total variation (TV)-based compressed sensing and statistical iterative reconstruction algorithms," *Physics in Medicine and Biology*, vol. 54, no. 19, pp. 5781–5804, Oct. 2009.

[84] Z. Tian, X. Jia, K. Yuan, T. Pan, and S. B. Jiang, "Low-dose CT reconstruction via edge-preserving total variation regularization," *Physics in Medicine and Biology*, vol. 56, no. 18, pp. 5949–5967, Sep. 2011.

[85] J. Provost and F. Lesage, "The Application of Compressed Sensing for Photo-Acoustic Tomography," *IEEE Transactions on Medical Imaging*, vol. 28, no. 4, pp. 585–594, Apr. 2009.

[86] Y. Lou, X. Zhang, S. Osher, and A. Bertozzi, "Image recovery via nonlocal operators," *Journal of Scientific Computing*, vol. 42, no. 2, pp. 185–197, 2010.

[87] G. Gilboa and S. Osher, "Nonlocal operators with applications to image processing," *Multiscale Modeling &amp; Simulation*, vol. 7, no. 3, pp. 1005–1028, 2008.

[88] J. Huang and F. Yang, "Compressed magnetic resonance imaging based on wavelet sparsity and nonlocal total variation," in *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2012, pp. 968–971.

[89] J. Liu, H. Ding, S. Molloi, X. Zhang, and H. Gao, "TICMR: Total Image Constrained Material Reconstruction via nonlocal total variation regularization for spectral CT," *IEEE Transactions on Medical Imaging*, pp. 1–1, 2016.

[90] X. Jia, Y. Lou, B. Dong, Z. Tian, and S. Jiang, "4d computed tomography reconstruction from few-projection data via temporal non-local regularization," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2010, pp. 143–150.

[91] M. Muja and D. G. Lowe, "Scalable nearest neighbor algorithms for high dimensional data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 11, pp. 2227–2240, 2014.

[92] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE signal processing magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[93] R. G. Baraniuk, "Compressive sensing," *IEEE signal processing magazine*, vol. 24, no. 4, 2007.

[94] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. E. Kelly, R. G. Baraniuk, and others, "Single-pixel imaging via compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, p. 83, 2008.

[95] R. Baraniuk, M. A. Davenport, M. F. Duarte, C. Hegde, and others, "An introduction to compressive sensing," *Connexions e-textbook*, 2011.

[96] J. Romberg, "Imaging via compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 14–20, 2008.

[97] N. Komodakis and J.-C. Pesquet, "Playing with Duality: An overview of recent primal-dual approaches for solving large-scale optimization problems," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 31–54, 2015.

[98] E. Esser, X. Zhang, and T. F. Chan, "A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science," *SIAM Journal on Imaging Sciences*, vol. 3, no. 4, pp. 1015–1046, 2010.

[99] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling &amp; Simulation*, vol. 4, no. 4, pp. 1168–1200, 2005.

[100] H. H. Bauschke and P. L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces.* Springer Science &amp; Business Media, 2011.

[101] N. Perraudin, J. Paratte, D. Shuman, V. Kalofolias, P. Vandergheynst, and D. K. Hammond, "GSPBOX: A toolbox for signal processing on graphs," *arXiv preprint arXiv:1408.5781*, 2014.

[102] T. Strohmer and R. Vershynin, "A randomized Kaczmarz algorithm with exponential convergence," *Journal of Fourier Analysis and Applications*, vol. 15, no. 2, pp. 262–278, 2009.

[103] F. Mahmood, L.-G. W. Öfverstedt, and B. U. Skoglund, *Extended field iterative reconstruction technique (efirt) for correlated noise removal.* USPTO, WIPO, JPO, Mar. 2014, uS Patent App. 14/770,245.

[104] A. Brandt, "Algebraic multigrid theory: The symmetric case," *Applied Mathematics and Computation*, vol. 19, no. 1-4, pp. 23–56, Jul. 1986.

[105] R. A. Crowther and A. Klug, "Three dimensional image reconstruction on an extended field—a fast, stable algorithm," *Nature*, vol. 251, no. 5475, pp. 490–492, Oct. 1974.

[106] A. Tikhonov, "Solution of incorrectly formulated problems and the regularization method," in *Soviet Math. Dokl.*, vol. 5, 1963, pp. 1035–1038.

[107] L. Bongini, D. Fanelli, F. Piazza, P. De Los Rios, S. Sandin, and U. Skoglund, "Freezing immunoglobulins to see them move," *Proceedings of the National Academy of Sciences*, vol. 101, no. 17, pp. 6466–6471, Apr. 2004.

[108] J. Wartiovaara, L.-G. Öfverstedt, J. Khoshnoodi, J. Zhang, E. Mäkelä, S. Sandin, V. Ruotsalainen, R. H. Cheng, H. Jalanko, U. Skoglund, and K. Tryggvason, "Nephrin strands contribute to a porous slit diaphragm scaffold as revealed by electron tomography," *Journal of Clinical Investigation*, vol. 114, no. 10, pp. 1475–1483, Nov. 2004.

[109] R. Akhouri, S. Goel, H. Furusho, U. Skoglund, and M. Wahlgren, "Architecture of Human IgM in Complex with P. falciparum Erythrocyte Membrane Protein 1," *Cell Reports*, vol. 14, no. 4, pp. 723–736, Feb. 2016.

[110] G. H. Golub, P. C. Hansen, and D. P. O'Leary, "Tikhonov regularization and total least squares," *SIAM Journal on Matrix Analysis and Applications*, vol. 21, no. 1, pp. 185–194, 1999.

[111] C. D. Meyer, *Matrix analysis and applied linear algebra.* Siam, 2000, vol. 2.

[112] Y. Censor and T. Elfving, "Block-iterative algorithms with diagonally scaled oblique projections for the linear feasibility problem," *SIAM Journal on Matrix Analysis and Applications*, vol. 24, no. 1, pp. 40–58, 2002.

[113] Y. Censor, T. Elfving, G. T. Herman, and T. Nikazad, "On diagonally relaxed orthogonal projection methods," *SIAM Journal on Scientific Computing*, vol. 30, no. 1, pp. 473–504, 2008.

[114] T. Elfving, T. Nikazad, and C. Popa, "A class of iterative methods: Semi-convergence, stopping rules, inconsistency, and constraining," *Linköpings Universitet Thesis*, 2010.

[115] T. Elfving, T. Nikazad, and P. C. Hansen, "Semi-convergence and relaxation parameters for a class of SIRT algorithms," *Electronic Transactions on Numerical Analysis*, vol. 37, pp. 321–336, 2010.

[116] L. Landweber, "An iteration formula for Fredholm integral equations of the first kind," *American journal of mathematics*, vol. 73, no. 3, pp. 615–624, 1951.

[117] T. Mizutani, K. Arai, M. Miyamoto, and Y. Kimura, "Application of silica-containing nano-composite emulsion to wall paint: A new environmentally safe paint of high performance," *Progress in Organic Coatings*, vol. 55, no. 3, pp. 276–283, Mar. 2006.

[118] F. Mahmood, M. Toots, L.-G. Öfverstedt, and U. Skoglund, "2d discrete fourier transform with simultaneous edge artifact removal for real-time applications." IEEE International Conference on Field Programmable Technology (FPT), Dec. 2015, pp. 236–239.

[119] J. A. Fessler and B. P. Sutton, "Nonuniform fast Fourier transforms using min-max interpolation," *IEEE Transactions on Signal Processing*, vol. 51, no. 2, pp. 560–574, 2003.

[120] R. Bracewell, “The fourier transform and its applications,” *New York*, vol. 5, 1965.

[121] L. R. Rabiner and B. Gold, “Theory and application of digital signal processing,” *Englewood Cliffs, NJ, Prentice-Hall, Inc., 1975. 777 p.*, vol. 1, 1975.

[122] L. Moisan, “Periodic plus smooth image decomposition,” *Journal of Mathematical Imaging and Vision*, vol. 39, no. 2, pp. 161–179, 2011.

[123] B. Akin, P. A. Milder, F. Franchetti, and J. C. Hoe, “Memory bandwidth efficient two-dimensional fast Fourier transform algorithm and implementation for large problem sizes,” in *Field-Programmable Custom Computing Machines (FCCM), 2012 IEEE 20th Annual International Symposium on.* IEEE, 2012, pp. 188–191.

[124] J. W. Cooley and J. W. Tukey, “An algorithm for the machine calculation of complex Fourier series,” *Mathematics of computation*, vol. 19, no. 90, pp. 297–301, 1965.

[125] H. Kee, N. Petersen, J. Kornerup, and S. S. Bhattacharyya, “Systematic generation of FPGA-based FFT implementations,” in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing.* IEEE, 2008, pp. 1413–1416.

[126] M. Frigo and S. G. Johnson, “FFTW: An adaptive software architecture for the FFT,” in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 3. IEEE, 1998, pp. 1381–1384.

[127] M. Puschel, J. M. Moura, J. R. Johnson, D. Padua, M. M. Veloso, B. W. Singer, J. Xiong, F. Franchetti, A. Gacic, Y. Voronenko, and others, “SPIRAL: Code generation for DSP transforms,” *Proceedings of the IEEE*, vol. 93, no. 2, pp. 232–275, 2005.

[128] E. Wang, Q. Zhang, B. Shen, G. Zhang, X. Lu, Q. Wu, and Y. Wang, “Intel math kernel library,” in *High-Performance Computing on the Intel® Xeon Phi^{TM}*. Springer, 2014, pp. 167–188.

[129] B. Akın, F. Franchetti, and J. C. Hoe, “FFTs with near-optimal memory access through block data layouts,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2014, pp. 3898–3902.

[130] B. Akin, F. Franchetti, and J. C. Hoe, “Understanding the design space of DRAM-optimized hardware FFT accelerators.” IEEE, Jun. 2014, pp. 248–255.

[131] C.-L. Yu, K. Irick, C. Chakrabarti, and V. Narayanan, “Multidimensional DFT IP generator for FPGA platforms,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 58, no. 4, pp. 755–764, 2011.

[132] D. He and Q. Sun, “A practical print-scan resilient watermarking scheme.” in *International Conference on Image Processing (1)*, 2005, pp. 257–260.

[133] D. G. Bailey, *Design for embedded image processing on FPGAs.* John Wiley & Sons, 2011.

[134] D. Bailey, “Implementing Machine Vision Systems Using FPGAs,” in *Machine Vision Handbook.* Springer, 2012, pp. 1103–1136.

[135] T. Lenart, M. Gustafsson, and V. Öwall, “A hardware acceleration platform for digital holographic imaging,” *Journal of Signal Processing Systems*, vol. 52, no. 3, pp. 297–311, 2008.

[136] G. H. Loh, “3d-stacked memory architectures for multi-core processors,” in *ACM SIGARCH computer architecture news*, vol. 36. IEEE Computer Society, 2008, pp. 453–464.

[137] Q. Zhu, B. Akin, H. E. Sumbul, F. Sadi, J. C. Hoe, L. Pileggi, and F. Franchetti, “A 3d-stacked logic-in-memory accelerator for application-specific data intensive computing,” in *3D Systems Integration Conference (3DIC), 2013 IEEE International.* IEEE, 2013, pp. 1–7.

[138] I. S. Uzun, A. Amira, and A. Bouridane, “FPGA implementations of fast Fourier transforms for real-time signal and image processing,” *IEE Proceedings-Vision, Image and Signal Processing*, vol. 152, no. 3, pp. 283–296, 2005.

[139] T. Dillon, “Two Virtex-II FPGAs deliver fastest, cheapest, best high-performance image processing system,” *Xilinx Xcell Journal*, vol. 41, pp. 70–73, 2001.

[140] A. Hast, “Robust and Invariant Phase Based Local Feature Matching,” in *Pattern Recognition (ICPR), 2014 22nd International Conference on.* IEEE, 2014, pp. 809–814.

[141] B. Galerne, Y. Gousseau, and J.-M. Morel, “Random phase textures: Theory and synthesis,” *IEEE Transactions on image processing*, vol. 20, no. 1, pp. 257–267, 2011.

[142] R. Hovden, Y. Jiang, H. L. Xin, and L. F. Kourkoutis, “Periodic Artifact Reduction in Fourier Transforms of Full Field Atomic Resolution Images,” *Microscopy and Microanalysis*, vol. 21, no. 02, pp. 436–441, 2015.

[143] D. G. Bailey, “The advantages and limitations of high level synthesis for FPGA based image processing,” in *Proceedings of the 9th International Conference on Distributed Smart Cameras.* ACM, 2015, pp. 134–139.

[144] W. Wang, B. Duan, C. Zhang, P. Zhang, and N. Sun, “Accelerating 2d FFT with non-power-of-two problem size on FPGA,” in *2010 International Conference on Reconfigurable Computing and FPGAs.* IEEE, 2010, pp. 208–213.

[145] F. Mahmood, N. Shahid, P. Vandergheynst, and U. Skoglund, “Graph-based sinogram denoising for tomographic reconstructions.” IEEE, Aug. 2016, pp. 3961–3664.

[146] N. Shahid, F. Grassi, and P. Vandergheynst, “Multilinear Low-Rank Tensors on Graphs & Applications,” *arXiv preprint arXiv:1611.04835*, 2016.

[147] A. Chambolle and T. Pock, “A first-order primal-dual algorithm for convex problems with applications to imaging,” *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, 2011.