SEGMENTING THE MALE PELVIC ORGANS FROM LIMITED ANGLE IMAGES WITH APPLICATION TO ART

Charles Brandon Frederick

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Biomedical Engineering.

Chapel Hill
2013

Approved by:

David Lalush

Stephen Pizer

Sha Chang

Wesley Snyder

Paul Segars

© 2013
Charles Brandon Frederick
ALL RIGHTS RESERVED

Abstract

CHARLES BRANDON FREDERICK: SEGMENTING THE MALE PELVIC ORGANS FROM LIMITED ANGLE IMAGES WITH APPLICATION TO ART.
(Under the direction of David Lalush and Stephen Pizer)

Prostate cancer is the second leading cause of cancer deaths in men, and external beam radiotherapy is a common method for treating prostate cancer. In a clinically state-of-the-art radiotherapy protocol, CT images are taken at treatment time and are used to properly position the patient with respect to the treatment device. In adaptive radiotherapy (ART), this image is used to approximate the actual radiation dose delivered to the patient and track the progress of therapy. Doing so, however, requires that the male pelvic organs of interest be segmented and that correspondence be established between the images (registration), such that cumulative delivered dose can be accumulated in a reference coordinate system. Because a typical prostate radiotherapy treatment is delivered over 30-40 daily fractions, there is a large non-therapeutic radiation dose delivered to the patient from daily imaging. In the interest of reducing this dose, gantry mounted limited angle imaging devices have been developed which reduce dose at the expense of image quality. However, in the male pelvis, such limited angle images are not suitable for the ART process using traditional methods.

In this work, a patient specific deformation model is developed that is sufficient for use with limited angle images. This model is learned from daily CT images taken during the first several treatment fractions. Limited angle imaging can then be used for the remaining fractions at decreased dose. When the parameters of this model are set, it provides segmentation of the prostate, bladder, and rectum, correspondence between the images, and a CT-like image that can be used for dose accumulation. However, intra-patient deformation in the male pelvis is complex and quality deformation models cannot be developed from a reasonable number of training images using traditional methods. This work solves this issue by partitioning the deformation to be explained into independent sub-models that explain deformation due to articulation, deformation near to the skin, deformation of the prostate bladder, and rectum, and any residual deformation. It is demonstrated that a model that segments the prostate with accuracy comparable to inter-expert variation can be developed from 16 daily images.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

ART adaptive radiotherapy

AP anterior-posterior

BCH Baker-Campbell-Hausdorff formula

BFGS Broyden-Fletcher-Goldfarb-Shanno

CBCT cone beam CT

CPU central processing unit

CT X-ray computed tomography

CUDA Compute Unified Device Architecture

DSC Dice's similarity coefficient

DVF displacement vector field

EBRT external beam radiotherapy

EM expectation-maximization

FBCT fan beam CT

FBP filtered backprojection

FIR finite impulse response

FOV field of view

FST Fourier Slice Theorem

GPL GNU General Public License

GPU graphics processing unit

GW-LNCC sum of Gaussian weighted local normalized cross correlation

HIFU high intensity focused ultrasound

IGRT image-guided radiotherapy

IIR infinite impulse response

IRRR iterative reprojection-reconstruction registration

LA-CBCT limited angle cone beam CT

LDDMM large deformation diffeomorphic metric mapping

LR left-right

ML maximum likelihood

MR magnetic resonance

NCC normalized cross correlation

NST nanotube stationary tomosynthesis

OAR organ at risk

ODE ordinary differential equation

PBR prostate-bladder-rectum organ complex

PCA principal component analysis

PDE partial differential equation

PET positron emission tomography

PGA principal geodesic analysis

PSA prostate-specific antigen

RCCT respiratory-correlated CT

SART simultaneous algebraic reconstruction technique

SBP simple backprojection

SI superior-inferior

S-rep skeletal representation

SSD sum of squared differences

VVF velocity vector field

WLS weighted least-squares

Chapter 1

Introduction

1.1 Motivation

1.1.1 Image Guided Radiotherapy

Radiotherapy is the targeted delivery of radiation to tissue. Typically, it is used as a component in the treatment of cancer with the goal of damaging cancerous or potentially cancerous tissue while minimizing damage to healthy tissue. It can be used as a complement to or replacement for potentially more invasive procedures such as surgical resection and chemotherapy. This document focuses on applications for prostate cancer using external beam radiotherapy (EBRT), where therapeutic radiation is provided by an external source. Accurate delivery and monitoring of this therapy requires precise knowledge of the patient's anatomy when the radiation is being delivered. This work intends to support the long term goal of providing this knowledge in the male pelvis solely from a planning image and limited angle imaging at treatment-time. Such an approach is faster than the current state of the art and has reduced non-therapeutic radiation dose.

In EBRT, a planning image, typically X-ray computed tomography (CT), is taken; the known or suspected tumor volume and radio-sensitive or nearby organs at risk (OARs) are segmented; and a prescription is constructed for delivering a certain *dose* of radiation to the diseased tissue. This planning image is used to construct a treatment plan, which consists of a set of beams and associated parameters for those beams that include position and direction and cross-sectional radiation intensity pattern. Ideally, this plan delivers the prescribed dose only to the diseased tissue with no radiation dose delivered to healthy tissue. However, radiation necessarily deposits energy into the tissue through which it passes, so the parameters of the plan must be optimized to minimize dose to OARs and spread stray dose among healthy tissues, minimizing *hot spots*. After the completion and verification of the treatment plan, the patient is manually aligned to the treatment device, and the plan is executed, typically in some number of daily *fractions*.

The accuracy of the delivered plan is dependent on the geometric configuration of the patient with respect to the coordinate system of the treatment device, the coordinate system where the treatment plan is constructed. This has two main aspects. First, the patient must be positioned accurately with respect to the treatment device, and, second, the configuration of non-rigid internal organs affects the actual distribution of dose delivered. In the male pelvis, the prostate itself remains relatively rigid, but changes in rectal and bladder contents between planning-time and treatment-time can move the prostate out of the high dose region intended by the plan and move the nearby radio-sensitive bladder or rectum into the high dose region, resulting in an under-dose to the intended treatment volume and an over-dose to the radio-sensitive structure. This can lead to undesired side effects. For example, radiation damage to the rectal wall can result in a condition known as radiation proctitis [34].

In order to increase the accuracy of plan delivery, an image can be taken to provide geometric information about the patient at treatment-time, leading to image-guided radiotherapy (IGRT). The modality of the image can vary based on the treatment site and available technology, ranging from typical 2D X-ray images and portal images taken with the radiation from the treatment device, to various optical imaging techniques, to fully tomographic modalities, such at CT and magnetic resonance (MR). The image is then used to verify and adjust patient position and can also be used to estimate the actual radiation dose delivered during the fraction for treatment monitoring and adjustment. This monitoring leads to the emerging clinical practice of adaptive radiotherapy (ART) [75]. With ART, under- and over-dosing can be tracked, and the treatment plan can be modified accordingly.

Patient anatomical changes, such as tumor shrinkage and weight loss, can also invalidate the original plan, requiring re-planning. In the male pelvis, daily anatomical changes can be large, meaning that to deliver a highly accurate treatment, it could be beneficial to construct a new daily treatment plan reflecting the patient's daily anatomical configuration. Such an online adaptive radiotherapy procedure is in an experimental state but could potentially become commonplace due to the ability to better conform the dose distribution to the target anatomy, reducing side effects[54, 48] and potentially decreasing the number of fractions required for treatment [74], thus decreasing cost.

To accomplish the goals of ART three pieces of information are required from each fraction:

1. Delineation of the treatment volume and nearby OARs, in order to correctly position the patient.

2. A CT-like image of the patient, used in estimating the actual dose distribution delivered during the fraction.

3. Correspondence between that patient's anatomy and some common coordinate system, most likely, the planning image, in order to accumulate the dose delivered in the fractions so far.

Taken together, this information can improve the accuracy of delivery by correcting patient positioning with respect to soft tissue, monitoring the progress of treatment for potential adjustment, and potentially correcting and re-optimizing the plan prior to the delivery of each fraction.

The increased delivery accuracy and therapy monitoring provided by adaptive IGRT can improve patient outcome, but these procedures seem to require a daily CT image for greatest benefit. Acquiring this image requires additional time, increased non-therapeutic imaging dose, and, often, the purchase of an expensive, dedicated imaging device. Risks of increased non-therapeutic dose include the potentially inducing secondary cancers and more acute side effects such as radiation eyrthema (reddening of the skin) [51]. Therefore, achieving similar outcomes with less non-therapeutic radiation dose is desirable. Because it uses fewer X-ray projections than CT imaging, *limited angle imaging*, X-ray imaging with an angular range less than sufficient to uniquely determine a fully 3D image, can be used to reduce non-therapeutic imaging dose. In addition to decreasing dose, limited angle imaging can decrease imaging time, but such benefits do not come without cost.

1.1.2 Limited Angle Imaging

In order to uniquely determine a 3D image from a set of 2D projection images, projection images must be acquired from a sufficient number and range of angles. Considerations regarding these criteria are discussed in section 2.2. In the event that these sampling criteria are not met, the image is said to be *limited angle*. When no additional information is available, limited angle imaging produces poor quality, artifact-ridden images that do not provide any of the information needed to perform ART. Since the missing information is fundamentally unmeasured, in order to obtain quality images, missing information must be inferred *a priori* from some outside source. The information actually required is some prior knowledge about the patient's spatial attenuation distribution at treatment-time. The patient's planning CT can approximate this. However, it has undergone some unknown transformation, and, for the attenuation distribution to be useful, this transformation must be recovered.

Since it has been assumed that some deformed version of the patient's planning CT provides the required data, the problem is recovering the transformation, making this a limited angle registration problem rather than a limited angle reconstruction problem, where a reconstruction problem is one where the solution is an image and a registration problem is one where the solution is a transformation. These two problems are dual to each other, and they both fail for the same reason. Given the information measured from limited angle images, there is not a unique solution to either the reconstruction or registration problem, and because the space of all possible solutions is much, much larger than the space of feasible solutions, the specific solution determined is not likely to be a desirable one. To solve this problem, a model can be developed that admits feasible solutions and excludes almost all infeasible solutions.

These types of models are typically learned with statistical dimensionality reduction methods, such as principal component analysis (PCA). The space of all diffeomorphisms (smooth, non-rigid transformations with smooth inverse), the type of transformation with which we are concerned, is infinite dimensional. If it is known *a priori* that the transformation that a patient undergoes is rigid, the dimensionality is reduced to 6, and limited angle rigid registration performs adequately. This work proposes to find a lower dimensional subspace of all diffeomorphisms that reflects feasible deformations for a particular patient and to solve the limited angle registration problem in that subspace. Such a subspace can be learned by examining daily CT images of a patient in order to develop an intra-patient deformation model.

1.1.3 Intra-patient Deformation Models

A first approach to the construction of a deformation model involves the collection of a set of 3D images that are to be registered together to obtain a single atlas image and deformations to each of the training images. The statistical method PCA can then be used to determine a lower dimensional representation of the space of likely transformations that the patient can undergo (a shape space). The parameters of this space can then be determined from limited angle images. This method is particularly successful in the thorax region, where most deformation is the result of respiration. The short duration and relatively periodic nature of the respiratory cycle allows the measurement of a respiratory-correlated CT (RCCT), effectively a set of images of the patient at several phases of the respiratory cycle. The RCCT can be measured once at planning-time and provides an adequate set of training images for the development of a PCA model of the respiratory cycle. Limited angle methods using PCA models like this have been developed in [16, 14, 45, 44].

PCA models for limited angle registration are particularly successful because of the relatively simple nature of respiratory motion and the presence of high-contrast surrogate structures. A PCA model in the lung can typically explain $> 95\%$ of the variation in a 10-phase RCCT with only 3 modes of variation, requiring only 3 parameters (plus 6 rigid parameters) to accurately determine a planning-time to treatment-time registration [14]. Additionally, the high-contrast lung boundaries, especially with the diaphragm, almost entirely indicate the respiratory phase of the subject. This, in addition to the few parameters to be determined, enables fast registration from even a single projection, effectively allowing real-time tumor tracking.

Many of these beneficial features are not found in the male pelvis. Primarily, the modes of variation are much more complex. Bladder and rectal contents change daily, are independent, and can induce large deformations. Even with a set of 16 training images, models similar to the PCA method developed for the thorax do not provide sufficient accuracy for clinical use. In addition, there are few high-contrast structures that help indicate the position of regions of interest, and the ones that do exist, namely, the skin, bones, and rectal gas, do not provide adequate information about the location of the prostate. This requires more projections over a greater angular range than in the thorax case in order to resolve the indicative low-contrast boundaries. Finally, the major challenge to the male pelvis is that major sources of deformation, bladder and rectal contents, occur on inter-fractional time scales and, thus, cannot be determined from an RCCT-like method.

This method and those similar to it are intra-patient deformation models, meaning that they describe the deformation within a single patient rather than among many. To enable the clinically ideal imaging protocol, where a single planning image is taken and only limited angle images are taken at treatment-time, an inter-patient model must be developed. Developing such a model would probably require summarizing information from multiple intra-patient models in such a way that summarized intra-patient models can be applied to a novel patient. Developing such a method is a complex undertaking and is beyond the scope of this work. Therefore, the first research question is if an adequate intra-patient model can be developed from existing clinical data. However, this method can be applied to a less clinically ideal protocol where the patient receives a daily CT for some number of initial fractions, and limited angle imaging is used the remaining fractions. This is discussed in chapter 5.

The main weakness of the methods described above derive from the nature of PCA. First, PCA discovers a single global shape space, which, based only on the nature of the observed data, can incorrectly imply that certain deformations are correlated, when they are not actually so. For example, if the patient happened to have a full bladder on days when the left leg was more abducted, this may be reflected in the shape space, and the position of the leg may improperly bias the estimation of the fullness of the bladder. However, the fullness of the bladder *does* affect the position of the prostate. Second, the number of modes of variation that can be discovered from a PCA model is limited to the number of images in the model (because that many modes absolutely describes all the variance observed in the population), and it is not usually desirable to use all the modes of variation from a PCA model (because that can lead to overfitting). If more data were available, it is possible that such simple PCA models could succeed. Third, PCA ignores anatomical constraints. Bones move rigidly, and, even if the registration method used produced deformations that were rigid in bony regions, a PCA will not reflect this fact.

The method developed here alleviates these problems by separating important and identifiable anatomical regions into several, independent, *a priori* correlated deformations. Such a technique divides a single global transformation into several local ones. This partitions the total variance to be explained in the complete model into several PCA models, increasing the specificity of each. Finally, it allows the use of a different deformation model which maintains rigidity in bony regions.

This method uses 4 specific transformations to provide adequate performance:

1. A transformation that uses high-contrast bony anatomy as an intermediate step between an initial rigid registration prior to the application of the deformation model. This transformation approximates rigid, articulated motion in the pelvis and femurs and its effects on nearby tissue.

2. A transformation that explains deformation of the skin and nearby tissue. The skin-air boundary has high contrast. This means that it contributes a large amount of signal to each of the projection images. Large inaccuracies in recovering the skin deformation can easily overwhelm the signal from the low contrast organs of interest.

3. A transformation that explains the deformations of the prostate, bladder, and rectum as an organ complex. This transformation explains the daily changes in bladder and rectal volume and how those changes effect the position and shape of the prostate.

4. A residual transformation that explains deformation not contained in any of the previous transformations. This mainly represents changes in muscle and fat over the course of therapy.

This dissertation shows that these transformations explain enough variation in the pelvic region of a male patient by registration to limited angle images to provide deformable segmentations of the prostate which are comparable in accuracy to inter-expert variation in CT segmentation.

Each of these transformations is non-Euclidean, meaning that standard mathematical operations must be modified to handle the non-flat nature of their spaces. To handle this issue, this work heavily employs the Log-Euclidean framework [2], which is a method of mapping the non-Euclidean space of transformations into a Euclidean one in which convenient mathematical operations can be developed. This mapping is known as the Logarithm of the transformation and is a generalization of the standard logarithm. Similarly, it is the inverse of the Exponential. Specifically, this work extends the poly-rigid transformation [3] (a manner of spatially combining multiple rigid transformations together) and the symmetric Log demons registration method (an extension of the Demons registration method to the Log-Euclidean Framework which employs properties of the Log domain representation to ensure that the transformation is smooth and has a smooth inverse and that the transformations determined when image A is registered to image B and when image B is registered to image A are inverses).

1.2 Thesis and Contributions

Thesis: Intra-patient motion models learned from daily CT images and using multiple, independent statistical deformation models can be used with limited angle projection images to accurately predict the position of the prostate, bladder, and rectum. The correspondence provided by this method allows the estimation of a CT-like image of sufficient quality to enable accumulation of dose from daily fractions into a common coordinate system for treatment monitoring and patient setup adjustment.

The contributions described in this dissertation are as follows:

1. The development and evaluation of a method that accounts for articulation of rigid structures, and the use of that method to reduce variation to be explained by successive transformations. This method rigidly transforms bony regions and approximates the effects that such bony transformations would have on surrounding tissues.

2. The development of an intra-subject deformation model in the male pelvis for use in registration via limited angle imaging that makes use of four models independently accounting variation due to articulation, skin deformation, deformation of the prostate, bladder, and rectum, and residual deformation not accounted for by the previous models. This multi-deformation model paradigm partitions the variance to be explained into several independent models to partially overcome the limitations of PCA.

3. A method to improve the convergence of symmetric, group-wise Log demons using the Log-Euclidean Fréchet mean of diffeomorphisms, where the Fréchet mean is a generalization of the Euclidean mean to non-Euclidean spaces.

4. An evaluation of the usefulness of the Log-Euclidean Framework for dimensionality reduction on diffeomorphisms in the above scenario.

5. An evaluation of the usefulness of the method developed by combining items 1-3 in limited angle registration for IGRT in the male pelvis.

6. A method for the masking of gas bubbles to increase the accuracy of atlases, registrations, and deformation models learned from registration in regions influenced by transient gas bubbles.

7. An evaluation of the number of daily CTs required for training data is performed. In a clinical application of this method, a patient would have daily CTs for the first several fractions to provide sufficient data to learn the models. The method could then be used at reduced dose for the remaining fractions.

1.3 Document Overview

This chapter summarizes the motivations and contributions of this dissertation. The remaining chapters are summarized as follows:

Chapter 2 provides necessary background knowledge for detailed understanding of the contributions of this dissertation, including IGRT, tomography, and a mathematical treatment of diffeomorphisms and registration.

Chapter 3 increases the reader's understanding of the overall construction of this method with a more detailed description of the transformations used in its development, their specific purposes, and extensions of existing methods.

Chapter 4 develops a poly-rigid method accounting for the articulated, rigid motion of bony tissue and its effects on surrounding soft tissue.

Chapter 5 combines the contributions from chapter 4 with non-rigid deformation models and applies the combined method to simulated clinical images.

Chapter 6 discusses the contributions made by this dissertation and possible future applications of this work.

Appendix A comments on the value of parallel graphics processing unit (GPU) computation in medical imaging and introduces the reader to NVidia's Compute Unified Device Architecture (CUDA) architecture.

Appendix B provides further details about the implementation of the Log-Euclidean framework with the goal of elucidating the abstract mathematical details.

Chapter 2

Background

2.1 Radiation Therapy in the Male Pelvis

2.1.1 Prostate Cancer and Treatment

The target area for methods in this dissertation is prostate cancer in the male pelvis. In 2011 in the United States, prostate cancer was the second most common cause of cancer deaths (behind lung) in men, and the most common cause of new cancer diagnoses over all [65]. This has created a high demand for prostate cancer treatment. However, there are controversies regarding the need for intervention in early-stage prostate cancer. The prostate-specific antigen (PSA) test is a non-invasive blood test, and its use as a non-invasive screening mechanism has led to an increase in diagnosis of early-stage, localized prostate cancer [18]. Many of these diagnoses may reflect slow-growing cancers and those that are unlikely to metastasize. Intervention for this type of disease may not be necessary for many years or may never be necessary. This potential over-diagnosis and over-treatment has led [50] to recommended against PSA for routine screening. In an attempt to reduce over-treatment, many physicians have adopted an active surveillance approach, where decisions on treatment are based on changes in screening results (e.g., digital exam, ultrasound, or PSA levels), or watchful waiting approach, where treatment follows from changes in symptoms [66], when deciding on clinical intervention following high PSA values. There is still no clear consensus on whether immediate treatment, active surveillance, or watchful waiting is the best approach to reducing mortality and increasing quality of life[18, 36]. Despite the above concerns, many patients will receive treatment for their prostate cancer each year.

There are three typical modes of cancer treatment: surgical, chemical, and radiation. Because cancer is a pernicious and resilient disease, a multi-faceted approach with many complementing treatments is often used. For example, chemotherapy is often combined with radiation. For disease entirely confined to the prostate, surgical prostatectomy, the complete removal of the prostate, can be a good option. However, prostatectomy can have life altering side effects such as incontinence and erectile dysfunction [72]. Aside from a traditional open surgery, this procedure is often laproscopic or robotically-assisted (for example, using the Da Vinci surgical robot [7]). However, it has not been shown that there is a significant difference in urological side effects between each of the procedures [23]. Chemotherapy with endocrine affecting substances [32] are also possible interventions. Hormonal therapies deplete circulating androgens (including testosterone), effectively chemical castration. This method is often preferable to orchidectomy, surgical removal of the testicles [32]. Most prostate cancer patients tolerate other forms of chemotherapy poorly, and most hormone responsive cancers eventually become androgen resistant [67]. Additional non-radiation therapies are high intensity focused ultrasound (HIFU), where ultrasound energy is focused on the prostate tissue to heat it and kill cancer cells [8], and cryosurgery, where the prostate is frozen [56].

Brachytherapy is another radiation-based therapy for prostate cancer. Brachytherapy is the permanent or temporary placement or implantation of a sealed radioactive source near to the treatment region [40]. There, these sources typically use low energy gamma rays irradiate tissue near to the seed. Implanted seed brachytherapy also benefits image-guided radiotherapy (IGRT) in that the radio-opaque seeds can be used as fiducials for localizing the prostate, which can change positions relative to the bony anatomy visible on projections images.

2.1.2 Radiation Therapy Process

IGRT divides into two distinct times, planning-time and treatment-time. At planning-time, a diagnostic quality 3D or 4D image of that patient is performed. This is usually a X-ray computed tomography (CT) but may include positron emission tomography (PET) or magnetic resonance (MR) images. On this image, structures of interest are identified and segmented. In the male pelvis, structures of interest may include the prostate, bladder, rectum, seminal vesicles, urethra, femoral heads, and skin. A physician prescribes a treatment plan, which includes the total dose to be delivered to the target and the number of fractions over which that dose will be delivered. A dosimetrist then uses a beam simulation and dose calculation program to develop a plan by adjusting beam parameters to fulfill the prescription while minimizing dose to healthy tissue and organs at risk (OARs). Beam parameters include number and angle of treatment beams, collimation (adjustment of the shape of the radiation pattern), energy, and intensity pattern. This treatment plan is then carried out at treatment-time.

At each treatment-time, the patient is positioned and the actual treatment is carried out. First the patient is manually positioned with the guidance of calibrated laser guides and external fiducial markers. Positioning may include custom restraints. Typically, some type of imaging is carried out to ensure that the patient is correctly positioned and possibly to observe whether inter-fractional anatomical changes are too large for the original plan to remain effective. The patient is then repositioned, if necessary. This may include simple adjustments of the treatment table, manual repositioning, or replanning. The treatment plan can then be carried out. Here speed of imaging, computation, and treatment is paramount because this directly translates into the cost and accuracy of procedures.

In delivering any medical procedure, costs must be considered. If the procedure cannot be made efficient and cost-effective, it will not be widely adopted. In IGRT, the cost is nearly equivalent to time. IGRT as it is widely practiced today is still requires many manual, time-consuming tasks be performed by expert individuals demanding high salaries [33]. This starts at planning-time, when segmentations and dosimetry are often performed with large amounts of manual intervention, although inroads are being made towards more fully automated segmentation [58] and faster dose calculations and treatment plan optimization [31]. Treatment-time is major target area for speed and cost improvements by increasing the speed of imaging and decreasing computation time. In addition to requiring highly trained technicians, treatment-time also occupies expensive treatment and imaging devices. Reducing treatment time means that more patients can be treated during the clinical day, thus reducing the marginal cost of treatment. The benefits to patients are not to be understated. During treatment, the patient must remain still in order for positioning to remain accurate, but involuntary motions, including respiratory and gastrointestinal motions, will occur. Effectively, the sooner treatment can be completed after imaging, the more accurate positioning is due to a lower likelihood of involuntary and voluntary motions affecting patient positioning. This imposes a *medical real-time* limit on the amount of computation and imaging that can be performed. A 10 minute MRI or registration procedure is not feasible unless great clinical utility is demonstrated.

2.1.3 Devices and Methods for Imaging and Treatment

While many varied therapeutic devices exist, a standard treatment device is a linear accelerator mounted on a rotating gantry. In the linear accelerator, electrons are accelerated to an energy on the order of 1-25 MeV incident on an anode from which X-rays are produced. The X-ray flux passes through hardening and flattening filters that filter out low energy X-rays, which are not necessarily desirable for therapy, and that selectively attenuate the beam so that the intensity is more uniform, respectively. The beam is then collimated using either a fixed or adjustable collimator. This shapes the beam to attenuate non-therapeutic radiation. In addition to photons, gamma rays (which are the same as X-ray photons but are produced by radioactive decay rather than by Bremsstrahlung and characteristic X-rays), electrons, protons, neutrons, or heavy ions can be used for therapy with unique costs and benefits.

Many treatment-time imaging solutions are available, and they can be divided into several categories. Non-ionizing methods, such as MR and ultrasound, and temporal 4D methods. MR methods could be useful to IGRT due to their good soft tissue contrast, but their low speed and requirement for a high strength magnet usually prohibits their use for image guidance at treatment-time. Ultrasound can have poor contrast and generally require manual acquisition which limits its utility for external beam radiation therapy, as well. Temporal methods are not discussed here to due the small influence of the respiratory cycle on the male pelvis. However, 4D cone beam methods in the presence of motion are equivalent to limited angle problems, and similar methods to this can be applied there [45]. Ionizing methods can be divided based on the energy of the photons: MV or kV, and the angular sampling: complete, cone beam, limited, or projective. The choice of energy is a trade-off between convenience and soft tissue contrast. Higher energy photons are produced natively by the treatment device and do not require a separate detector but result in inherently lower soft tissue contrast.

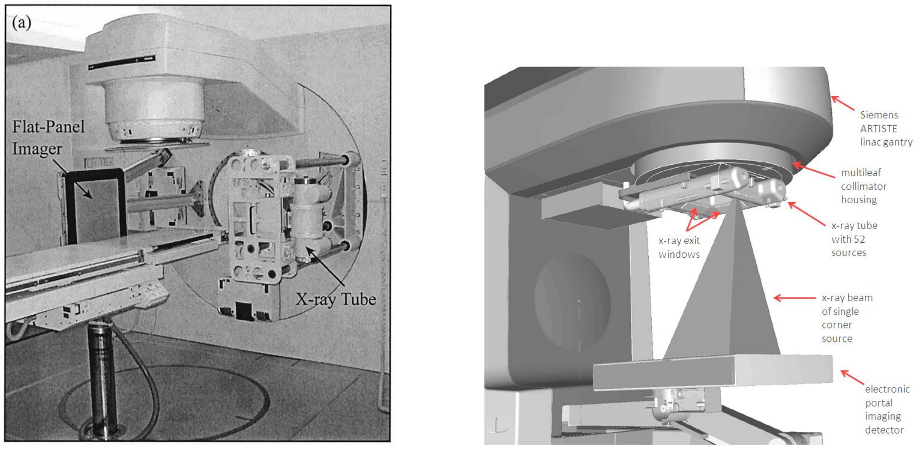
Production of lower energy photons may require an additional source and possibly an additional detector. However, the hardening filter in the MV beam path could be removed to increase the flux of low energy photons, increasing contrast. Kilovoltage X-rays are usually preferred to provide images for IGRT.

Ideally, image guidance would be provided by a diagnostic kV fan beam CT (FBCT). This is the only complete angular sampling system commonly used at treatment-time. Non-helical cone beam imagers, while they satisfy 2D angular sampling requirements, do not satisfy an additional sampling requirement, the Tuy condition [68], and have additional detriments, so they are addressed separately. An FBCT, however, is geometrically incompatible with a gantry system. As such, it must be implemented separately and requires the purchase of an additional independent diagnostic system. In such a system, the positioned patient is translated out of the treatment field of view (FOV) into that of the CT system. This can require additional imaging time (although an FBCT itself is very fast), disallows simultaneous imaging and treatment for monitoring purposes, and may induce deformation in the subject. A cone beam CT (CBCT) system can be built integrated with the rotating gantry and use either MV or kV radiation but can take a significant amount of time due to the slow rotational speed of the large, heavy gantry which is necessary for circular CBCT completeness. In regions influenced by respiratory motion, this motion is blurred due to the long acquisition times unless the acquisition is gated. This either introduces a significant amount of additional imaging time (because a stop-and-wait approach is necessary) or becomes a limited angular sampling problem (due to division of projections into respiratory bins). Fan beam and cone beam systems also have a high non-therapeutic imaging dose, which is compounded by daily imaging. Single or orthogonal projective imaging is the most traditional method. It is fast and low dose but provides hardly any soft tissue contrast. However, it can be suitable for regions that have high contrast and rigidity, such as the brain.

This work is focused on limited angle imaging geometries where X-ray projections are gathered over an incomplete arc rather than a complete arc as in cone beam or complete geometries. These reduce the dose and time over fan beam and cone beam systems, sacrificing image quality, contrast, and resolution, while gaining image quality, contrast, and resolution over projective imaging at the cost of dose and, possibly, time. Limited angle geometries demand more novel image analysis techniques over fan beam or cone beam methods but exist along a continuum where methods with nearly complete angular sampling can be treated similarly to 3D methods and those with nearly projective angular sampling can be treated similarly to 2D projective methods. The limited angle problem is discussed further in later sections, including section 2.2.4.

There are two similar techniques for obtaining limited angle images with gantry mounted kV imagers. The older makes use of a single rotating kV X-ray source and detector pair mounted orthogonal to the treatment direction. Such a setup can be used to acquire CBCTs [37], but, with rotation over a limited arc, it can also be used to acquire limited angle images [29]. This method is known as limited angle cone beam CT (LA-CBCT). A more novel solution makes use of multiple independent kV sources mounted about the beam portal such that the MV treatment detector can be used for imaging [47]. This method is known as nanotube stationary tomosynthesis (NST), owing to the use of carbon nanotubes as the electron source in the initial prototype and where tomosynthesis is a traditional name for limited angle imaging. Examples of both LA-CBCT and NST systems are shown in figure 2.1.

In addition to the limited angle problem, NST and LA-CBCT typically suffer from the truncation problem where the anatomical region being imaged is too large to be entirely projected on the detector. This loss of information results in artifacts in the reconstruction. This is a larger problem for NST owing to the stationary detector. The


(a)
Flat-Panel Imager
X-ray Tube
Siemens ARTISTE linac gantry
multileaf collimator housing
x-ray tube with 52 sources
x-ray exit windows
x-ray beam of single corner source
electronic portal imaging detector

(a) CBCT Device [37]

(b) NST Device [47]

Figure 2.1: Two gantry mounted kV imaging devices which can acquire limited angle images. a shows a CBCT/LA-CBCT source-detector pair mounted orthogonally to the treatment beam. Several projection images are taken over a variable angular range as the source-detector pair are rotated. b shows an NST system which has fixed angular range but does not require motion or a separate detector.

stationary detector and fixed sources also only allow a fixed angular coverage without gantry rotation, while the angular coverage of LA-CBCT is variable and can be adjusted for different requirements. Limited angle images suffer from an anisotropic resolution, having a direction with much poorer resolution than those orthogonal to it. LA-CBCT has the problem that the low resolution direction is orthogonal to the beam projection. If the treatment direction is anterior-posterior (AP), then the direction of poorest resolution is the left-right (LR) direction, due to imaging primarily in the LR direction, and positioning errors are most likely to appear in this direction. NST does not have this problem. If the imaging direction is AP, then errors are most likely to appear in the AP direction. Since the beam is transmitted through that tissue, dose is already being deposited anterior and posterior to the target region. Finally, since its sources are electronically addressable and it does not require motion, the imaging speed of NST is significantly faster, limited only by the readout speed of the flat panel detector. Both of these geometries have great clinical potential. The methods developed here are geometry agnostic and, as such, can be applied to either and many other possible variants.

2.2 Tomography

Tomography is the process of determining an image of some physical property of matter by probing the object from multiple angles with some type of wave or particle. While many types of tomography exist, this dissertation focuses on X-ray transmission tomography. In X-ray tomography, an X-ray source is incident on the matter of interest. The matter attenuates some of the photons while others are transmitted to a detector. The number of incident photons is compared with the number of transmitted photons to obtain a set of integral attenuation values. This is in contrast with emission tomography, where a radio-tracer is injected into the subject and emitted photons are used to obtain a distribution of radio-tracer within the patient. With a proper set of line integrals, an exact reconstruction can be determined. The following sections provide an introduction to tomography and its limits with application to this work.

2.2.1 Fourier Slice Theorem

The standard X-ray equation is

$$I\left(\boldsymbol{u}\right)=\int_{0}^{\infty} D\left(E\right) I_{0}\left(\boldsymbol{u},E\right) e^{-\int \mu(\boldsymbol{x},E)\ell(\boldsymbol{x},\boldsymbol{u})\,d\boldsymbol{x}}\,dE \tag{2.1}$$

where $\boldsymbol{I}\left(\boldsymbol{u}\right)$ is the detected intensity at detector location $\boldsymbol{u}$, $\boldsymbol{I}_0\left(\boldsymbol{u},E\right)$ is the incident photon intensity spectrum as a function of E, $D\left(E\right)$ is the detector energy sensitivity, $\mu\left(\boldsymbol{x},E\right)$ is the energy-dependent, spatial attenuation function to be obtained from tomography, and $\ell\left(\boldsymbol{x},\boldsymbol{u}\right)$ is a sampling function for $\boldsymbol{u}$, which is a line in practice. For the scope of this work, the energy dependence of this problem is intractable and is not considered, except to note that it does have measurable, non-negligible effects. With appropriate redefinition and approximation,

$$I(\boldsymbol{u}) = I_0(\boldsymbol{u})\, e^{-\int \mu(\boldsymbol{x})\ell(\boldsymbol{x},\boldsymbol{u})\; d\boldsymbol{x}} \tag{2.2}$$

One can then linearize the problem,

$$-\log \frac{I(u)}{I_0(\boldsymbol{u})} = p(\boldsymbol{u}) = \int \mu(\boldsymbol{x})\, \ell(\boldsymbol{x},\boldsymbol{u})\; d\boldsymbol{x} \tag{2.3}$$

where p are the processed projections.

For demonstration, one can define a 2D parallel beam geometry. This is in contrast to the fan beam and cone beam geometries implemented in modern CT. These geometries are shown in figure 2.2. The projection equation (2.3) can then be rewritten

$$\boldsymbol{u} = \boldsymbol{R}\boldsymbol{x} \tag{2.4}$$

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{2.5}$$

$$p_\theta(u) = \int_{-\infty}^{\infty} \mu(u,v)\; dv \tag{2.6}$$

where $\boldsymbol{R}$ is a rotation matrix applied to the object function rotating it such that the desired projection direction aligns with v. Equation (2.6) is the Radon transform, which transforms an image into a set of parallel line integrals taken at angle θ.

The FST relates the Radon transform at a particular θ to a line in the $2D$ Fourier domain of the original image. It is useful for understanding the sampling requirements to reconstruct a unique image and in developing an inversion method for the Radon transform. To demonstrate the FST, one takes the Fourier Transform of the projection and substitutes (2.6) acknowledging the relation $\boldsymbol{u} = \boldsymbol{R}\boldsymbol{x}$ and changing the variables

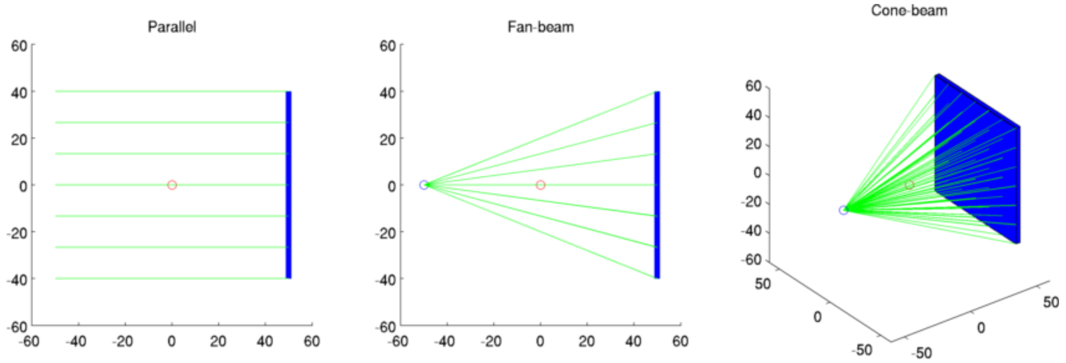


Figure 2.2: On the left, an example parallel beam geometry used to demonstrate the FST. In the center, a fan beam geometry similar to those used in diagnostic CT (except that clinical CT scanners use curved detectors). On the right, a cone beam geometry. This is common in radiation oncology, industrial, small animal, and research systems.

of integration.

$$P_\theta(\omega_u) = \int_{-\infty}^{\infty} p_{\boldsymbol{R}}(u, \ldots) e^{-j2\pi\omega_u u} \, du \tag{2.7}$$

$$= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} \mu(u, v) \, dv \right] e^{-j2\pi\omega_u u} \, du \tag{2.8}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mu(x, y) e^{-j2\pi\omega_u (x\cos\theta + y\sin\theta)} \, dx \, dy \tag{2.9}$$

The right hand side of equation (2.9) is the 2D Fourier transform of μ evaluated at the line in frequency space $u = \omega_u \cos\theta$ and $v = \omega_u \sin\theta$. So the 1D Fourier transform of a parallel projection of a function taken in the direction at angle θ from the y axis is equal to the 2D Fourier transform of the function evaluated at a line at angle θ from the x axis. This is shown graphically in figure 2.3.

The FST suggests that a function can be recovered from its projections by taking a sufficient number of projection samples, Fourier transforming, placing the transformed projections in the frequency domain, and inverting the Fourier transform. In practice, this is not the method used due to the requirements for interpolation. The actual methods used will be discussed in the following sections. A major use of the FST is to

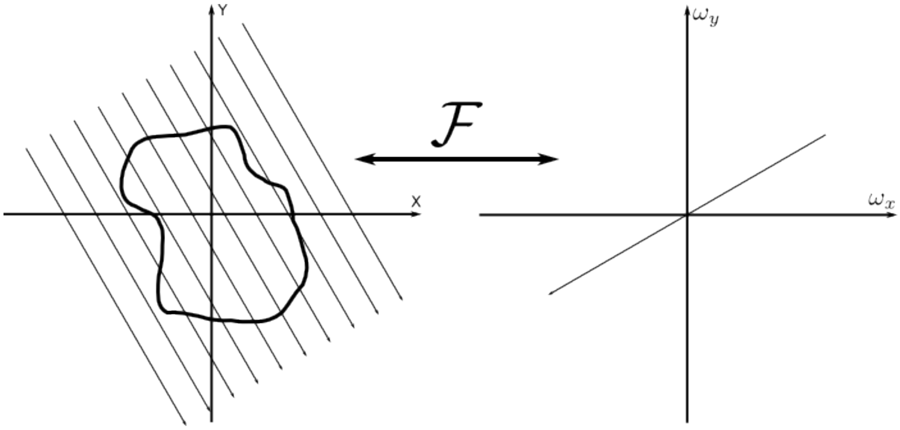


Figure 2.3: The FST relates the Fourier transform of a projection of a function to a radial line in the frequency domain representation of the function. The 1D Fourier transform of a projection on μ in the direction indicated by the rays on the left contains the frequencies of the 2D Fourier transform of μ in the direction orthogonal to the projection direction as indicated on the right.

demonstrate the sampling requirements for unique reconstruction. It is obvious that in order to completely sample the frequency domain for a parallel beam geometry one must sample over at least 180° (180°+ total fan angle for fan beam), and, due to the radial sampling of Fourier space, the higher frequencies are less densely sampled. In order to reconstruct higher spatial frequencies, more projection samples are required.

In this work, the projections are provided by devices that are incapable of completely sampling Fourier space due to mechanical constraints. The reconstructed images are necessarily under-sampled and will not correctly reflect the attenuation distribution of the object of interest unless additional information is provided. This limited angle problem is the motivation for this dissertation and will be discussed further in a later section.

2.2.2 Filtered Backprojection

Aside from the projection operator, the second fundamental building block of reconstruction is the backprojection operator. Since the projection operator is compressing

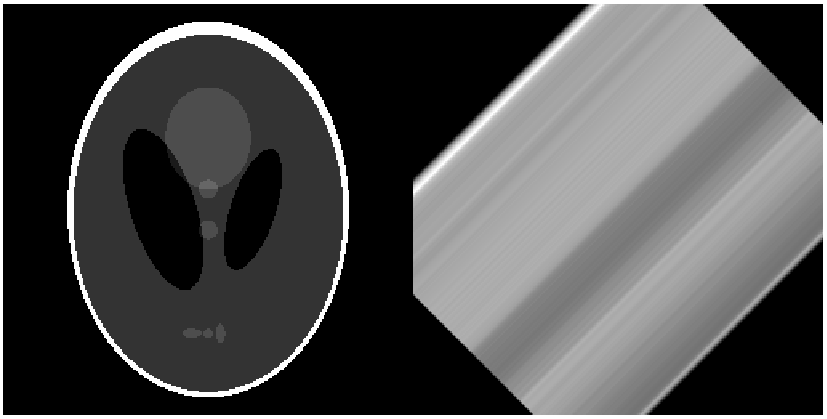


Figure 2.4: The modified Shepp-Logan head phantom and a single backprojection from a projection generated from a projection in the direction 45° counter-clockwise from the x axis

the image dimension by 1 along a set of rays, the backprojection operator is then taking that projection and spreading the accumulated intensity out along those same rays. An example backprojection is shown in figure 2.4. The simple backprojection (SBP) method obtains a reconstruction by summing the backprojections from all the projections. An example of simple backprojection is shown in figure 2.6. One can see that this is not the inverse of the projection operator. Rather, it is the adjoint. For a matrix operator, this is equivalent to multiplying by the transpose [5].

From the SBP example, one observes that the SBP reconstruction lacks the high frequency detail present in the original image. This comes from the FST in that low spatial frequencies are more densely sampled than high spatial frequencies. The sampling of Fourier space by a set of parallel projections is shown in figure 2.5, and, even if infinitely many projections were available, SBP would not be the correct answer. The solution to this issue is to weight the higher frequencies more strongly than the lower frequencies by applying a high pass filter to the projections before backprojection. This is the idea behind filtered backprojection (FBP).

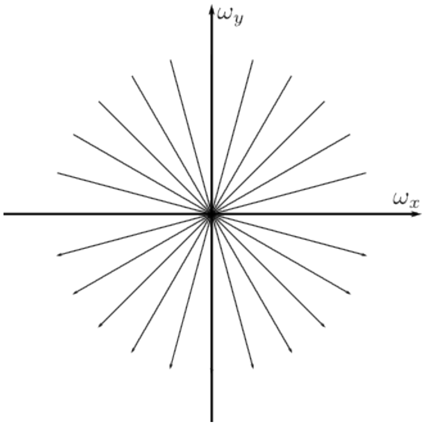


Figure 2.5: Because the low spatial frequencies are more densely sampled by projections, those frequencies are more strongly represented in the SBP reconstruction. This limitation is overcome by FBP.

To derive FBP, one starts with the 2D inverse Fourier transform of an object function M

$$\mu(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} M(\omega_x,\omega_y)\, e^{j2\pi(\omega_x x+\omega_y y)}\, d\omega_x\, d\omega_y \tag{2.10}$$

Then converting to polar coordinates with

$$\omega_x = r\cos\theta \tag{2.11}$$

$$\omega_y = r\sin\theta \tag{2.12}$$

$$d\omega_x\, d\omega_y = r\, dr\, d\theta \tag{2.13}$$

Then,

$$\mu(x,y) = \int_{0}^{2\pi}\int_{0}^{\infty} rM(r,\theta)\, e^{j2\pi r(x\cos\theta+y\sin\theta)}\, dr\, d\theta \tag{2.14}$$

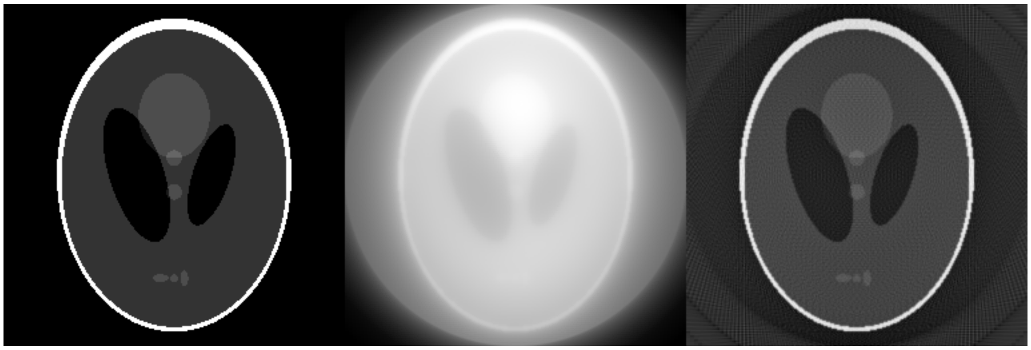


Figure 2.6: The modified Shepp-Logan head phantom and its reconstructions from 256 views over 360° using simple backprojection and filtered backprojection.

Since $M(r, \theta + \pi) = M(-r, \theta)$

$$\mu(x, y) = \int_0^\pi \left[\int_{-\infty}^\infty |r| \, M(r, \theta) \, e^{j2\pi r (x \cos\theta + y \sin\theta)} \, dr \right] d\theta \tag{2.15}$$

$P_\theta(r)$ can substituted because of the FST and $r = \omega_u$ at angle θ.

$$\mu(x, y) = \int_0^\pi \left[\int_{-\infty}^\infty |r| \, P_\theta(r) \, e^{j2\pi r (x \cos\theta + y \sin\theta)} \, dr \right] d\theta \tag{2.16}$$

Equation 2.16 provides an expression for exact reconstruction of μ in terms of continuous parallel projections over 180°. In order to reconstruct an object from its projections, those projections must be filtered with a kernel with frequency response $|\omega|$, a ramp filter, to properly invert the low pass characteristic of SBP. A sample FBP reconstruction is shown in figure 2.6. The artifacts in the reconstruction are a result of discretization and finite domains. To summarize the implementation, measured projections are filtered, here in the frequency domain, with a high-pass filter with response $|\omega|$ and then backprojected.

Equation 2.16 states an exact solution with continuous, noise free measurements. In practice, one can only measure a finite number of projections, meaning that some regions of the frequency domain will not be measured. Therefore, exact recovery is

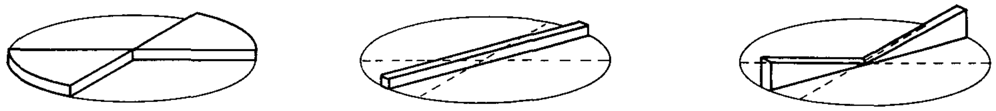


Figure 2.7: On the left is the spatial frequency sampling pattern required to completely sample the Fourier space of the image. In the center is the actual sampling pattern as provided by the FST. On the right is the approximation made by FBP. One can see that this is exact with continuous functions and infinite projections. In the finite projection case, some missing information remains. Image taken from [39].

not possible. Figure 2.7 shows the difference between the continuous situation and the discrete version. A more challenging issue for FBP is that any actual measurements will contain noise. This noise will be amplified by the high pass filter. Some problems with noise can be dealt with by apodizing (decreasing the amplification of high frequency components) the ramp filter to reduce the amount of high frequency noise. This comes at the cost of resolution in the reconstruction because some of the high frequency information from the projections is not used. While FBP is still used clinically due to its speed and familiarity, iterative reconstruction methods have been developed that can improve reconstruction quality by better handling noise without sacrificing resolution.

2.2.3 Iterative Methods

In demonstrating iterative methods, a notation change is made for convenience. Section 2.2.2 explained 2D parallel beam versions of the projection and backprojection operators necessary for reconstruction in terms of the integral. In modern CT scanners, the imaging geometry is fan beam or cone beam where an X-ray point source casts rays through an object incident on a 1D or 2D detector, respectively, which are rotated in a circular orbit. A great many other geometries are also possible, and it can be complicated to express these in a mathematically tractable manner. For iterative reconstruction, it can be preferable to express these operators without explicitly defining

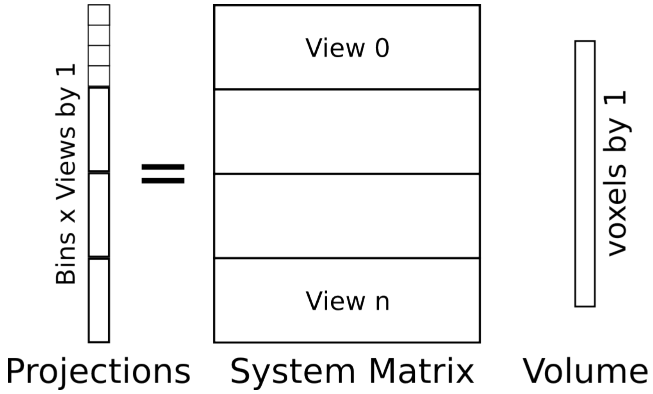


Figure 2.8: The matrix for is a convenient method for geometry independent analysis of iterative reconstruction algorithms. The interpretation of matrix entry C_{ij} is the contribution of image voxel j to projection bin j.

the geometry.

Since the integral operations are linear and all data are discrete, a matrix notation is introduced. Pixels or voxels of an image to be reconstructed and detector bins are regarded as a vector. The system matrix, $\boldsymbol{C}$, relates voxels in the image to detector bins in the projection. An example ordering is provided in 2.8. So if the image is an $N \times 1$ vector and the projection is $LM \times 1$ vector, then the system matrix is $LM \times N$, where N is the number of voxels in the image, L is the number of projections, and M is the number of detector bins. The interpretation of this matrix is that C_{ji} is the amount that voxel i contributes to detect bin and view j. In an example CBCT acquisition, C may be impossibly large. In a standard clinical acquisition, N may be on the order of 512^3, M may be 1024^2, and L may be 512, with $\boldsymbol{C}$ having 2^{56} elements. Such a matrix is impractical to store, even though it would be very sparse; thus, its elements are usually computed on the fly based on models of the geometry and physics of the true system.

FBP is a convolution-backprojection algorithm. Using this notation, it can be written $\hat{\boldsymbol{x}} = \boldsymbol{C}^T\boldsymbol{H}\boldsymbol{p}$, with measured projections $\boldsymbol{p}$ and a matrix representation of the high-pass filtering $\boldsymbol{H}$. This notation suggests a different solution procedure, solving a system of linear equations:

$$\boldsymbol{p} = \boldsymbol{C}\hat{\boldsymbol{x}} \tag{2.17}$$

$$\hat{\boldsymbol{x}} = \boldsymbol{C}^{\dagger}\boldsymbol{p} \tag{2.18}$$

One can imagine inverting this matrix once and then reconstructing exactly with a single matrix-vector multiplication. There are a number of issues with this approach. In general, this matrix is not square and may not have a solution. Therefore the previous equation employs the pseudoinverse, which gives the least-squares solution. $\boldsymbol{C}^{\dagger} = \left(\boldsymbol{C}^{T}\boldsymbol{C}\right)^{-1}\boldsymbol{C}^{T}$. This still requires the computing inverse of a large matrix, for which the naive algorithm requires $O\left(N^{3}\right)$ operations with $N = 2^{29}$, and would require 262144 TiB of storage given the previous geometry. Even though $\boldsymbol{C}$ is sparse, $\boldsymbol{C}^{\dagger}$ is not typically sparse. Such a computation will be impossible for many years to come. Furthermore, system matrices usually have high condition numbers. The condition number is a property of a matrix, independent of any algorithmic or floating point considerations, which roughly reflects the speed at which $\hat{\boldsymbol{x}}$ changes relative to changes in $\boldsymbol{p}$. So noise and small errors in $\boldsymbol{p}$ may result in large errors in $\hat{\boldsymbol{x}}$. Finally, $\boldsymbol{C}$ is always approximate, and errors in it exacerbate the previous statements. That being said, in a certain sense much of reconstruction methodologies can be imagined as solving these issues.

The goal of reconstruction is to find a image $\hat{\boldsymbol{x}}$ that best matches $\boldsymbol{p}$ given $\boldsymbol{C}$. So

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} d\left(\boldsymbol{p}, \boldsymbol{C}\boldsymbol{x}\right) \tag{2.19}$$

Iterative methods all share the property that they begin with an initial estimate for $\hat{\boldsymbol{x}}$ and try to improve that estimate iteratively by comparing the estimated projections

$\boldsymbol{C}\hat{\boldsymbol{x}}$ with $\boldsymbol{p}$. The initial estimate is first projected resulting in a set of estimated projections. The estimated projections are then compared with the measured projections, resulting in the projection space error. This error is then backprojected to get a reconstruction space error that is used to update the initial estimate. The process continues until a fixed number of iterations have elapsed or a convergence criterion is met. A graphical example of this flow is shown in figure 2.9. Iterative methods are sub-divided into algebraic methods and statistical methods, but there is not necessarily a distinct separation between the two. Algebraic methods attempt to solve the linear algebra problem in the previous paragraph. These are not discussed here since they are not highly relevant to this research. The class of interest here is statistical methods. These are often stated in a probabilistic framework.

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} \mathrm{P}\left[\boldsymbol{p}|\boldsymbol{x}\right] \tag{2.20}$$

The methods involve developing a probability distribution $\mathrm{P}\left[\boldsymbol{p}|\boldsymbol{x}\right]$ and then solving (2.20). This is commonly referred to as the maximum likelihood (ML) method.

As an introduction to these methods the weighted least-squares (WLS) method will be derived. Least-squares was mentioned above, and, due to its simplicity, it may be regarded as both an algebraic and ML method since it has both a statistical and linear algebraic interpretation. The WLS method is similar to least-squares and has an analytical solution, except that it contains an additional weighting matrix. By itself, WLS is not desirable for reconstruction, but, with modifications to ensure that solutions remain smooth, it is commonly used in CT reconstruction. In deriving WLS, a probability distribution is imposed on $\boldsymbol{p}$. In WLS, this is the multivariate Gaussian:

$$\mathrm{P}\left[\boldsymbol{p}|\boldsymbol{x}\right] = ke^{-\frac{1}{2}(\boldsymbol{p}-\boldsymbol{C}\boldsymbol{x})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{p}-\boldsymbol{C}\boldsymbol{x})} \tag{2.21}$$

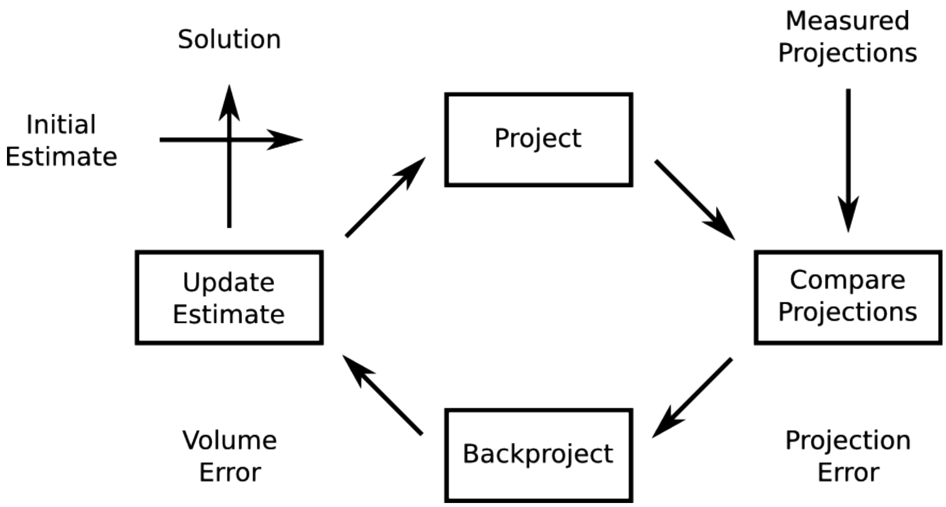


Figure 2.9: Iterative methods can be placed into this general framework where, given and initial estimate and measured projections, the algorithm iterates through the process of projecting the current estimate and comparing the projected estimate with the measured projections. This results in a projection space error that is then backprojected into the reconstruction domain, resulting in the volume space error. This volume space error is then used to update the current estimate. When the projection space error is sufficiently small or other convergence criteria are met, the current estimate is considered the solution.

where k is a constant which will become irrelevant. $\boldsymbol{p}$ is a random variable drawn from a Gaussian distribution with mean $\boldsymbol{Cx}$ and covariance $\boldsymbol{\Sigma}$. This matrix describes the covariance between pixels. It is typically taken to be diagonal indicating that each projection bin is independent and has a variance that is, for example, the Gaussian approximation to the Poisson statistics that govern discrete X-ray photons. Taking the logarithm of both sides and the derivative with respect to $\boldsymbol{x}$:

$$\log \mathrm{P}\left[\boldsymbol{p}|\boldsymbol{x}\right] = R\left(\boldsymbol{x}\right) = -\frac{1}{2}\left(\boldsymbol{p}-\boldsymbol{Cx}\right)^{T}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{p}-\boldsymbol{Cx}\right)+\log k \tag{2.22}$$

$$\frac{\partial}{\partial \boldsymbol{x}}R\left(\boldsymbol{x}\right) = \boldsymbol{C}^{T}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{p}-\boldsymbol{Cx}\right) \tag{2.23}$$

Then, setting $\frac{\partial}{\partial \boldsymbol{x}}R\left(\boldsymbol{x}\right)=0$ leads to the analytical solution, provided the inverse exists.

$$\boldsymbol{C}^{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{p} = \boldsymbol{C}^{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{Cx} \tag{2.24}$$

$$\boldsymbol{x} = \left[\boldsymbol{C}^{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{C}\right]^{-1}\boldsymbol{C}^{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{p} \tag{2.25}$$

For the reasons described above, the analytical solution is not evaluated. Instead, an optimization method is used which examines the local gradient of (2.22) to minimize the cost function. These methods first choose a step direction $\boldsymbol{h}$ and then determine the optimal step size t for that direction. Plugging in the step direction and setting $\frac{\partial}{\partial t}R\left(\boldsymbol{x}+t\boldsymbol{h}\right)=0$

$$\boldsymbol{x}^{i+1} = \boldsymbol{x}^{i}+t\boldsymbol{h} \tag{2.26}$$

$$R\left(\boldsymbol{x}+t\boldsymbol{h}\right) = -\frac{1}{2}\left(\boldsymbol{p}-\boldsymbol{Cx}-t\boldsymbol{Ch}\right)^{T}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{p}-\boldsymbol{Cx}-t\boldsymbol{Ch}\right) \tag{2.27}$$

$$= R\left(\boldsymbol{x}\right)+t\boldsymbol{h}^{T}\boldsymbol{C}^{T}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{p}-\boldsymbol{Cx}\right)-\frac{1}{2}t^{2}\boldsymbol{h}^{T}\boldsymbol{C}^{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{Ch} \tag{2.28}$$

$$\frac{\partial}{\partial t}R\left(\boldsymbol{x}+t\boldsymbol{h}\right) = \boldsymbol{h}^{T}\boldsymbol{C}^{T}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{p}-\boldsymbol{Cx}\right)-t\boldsymbol{h}^{T}\boldsymbol{C}^{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{Ch}=0 \tag{2.29}$$

$$t = \frac{\boldsymbol{h}^{T}\boldsymbol{C}^{T}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{p}-\boldsymbol{Cx}\right)}{\boldsymbol{h}^{T}\boldsymbol{C}^{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{Ch}} \tag{2.30}$$

The simplest choice is the method of steepest ascent (descent) that takes steps in the (opposite) direction of the gradient. However, this formula is valid for any $\boldsymbol{h}$, for which many choices are possible.

The main benefits of iterative methods are their ability to handle irregular geometries, their robustness to noise, and their ability to incorporate prior knowledge about the problem. The notation developed above allows iterative algorithms to be quickly adapted into geometries where it may be difficult to develop a proper reconstruction filter. In order to evaluate iterative algorithms in a novel imaging geometry, all that is necessary is to implement a function that multiplies a vector by the system matrix and another function for its transpose. Due to their ability to properly determine the probability of an image given knowledge of the imaging physics or the ability to better solve the linear system, iterative algorithms can find a more correct solution with less noise and fewer artifacts than a convolution-backprojection algorithm. This is especially useful in that it can be combined with prior knowledge, for example, that smoother images are more likely as reconstructions than rougher solutions.

Despite their obvious benefits, the clinical community has been slow to adopt iterative methods for CT reconstruction. The major issue is the great computational complexity. The major cost in reconstruction is projection and backprojection. The comparison and update stages are negligible in comparison. While FBP requires only convolution and backprojection, iterative algorithms require at least one projection and one backprojection per iteration, and 5-10 iterations may be required for adequate convergence. Since FBP produces adequate results and diagnostic radiology can require results quickly, FBP remains the primary choice in reconstruction algorithms. Iterative algorithms, however, have long been used in emission tomography where noise is greater and the much smaller problem size compared with CT makes their use more feasible and necessary. Recent advances in computing hardware, particularly graphics processing unit (GPU) accelerated computing, discussed in appendix A, have reduced the time and cost requirements to the point that real-time iterative reconstruction is possible, so iterative algorithms for transmission tomography will likely begin making appearances in the clinic.

2.2.4 Limited Angle Problem

Section 2.2.1 discussed the FST and the angular sampling requirements for unique reconstruction. If angular samples from a parallel geometry are acquired over 180°, then the frequency space of an object has been completely sampled and a unique reconstruction exists. If there are missing angles, the corresponding spatial frequencies simply have not been measured. The exact values cannot be recovered from the measured data alone. Coping with this missing information is the limited angle problem. This is not to say that having complete information is always necessary. Projection radiography is an extreme case of missing information, yet it retains high clinical utility. Limited angle X-ray tomography can be a valid diagnostic technique on its own. This method is traditionally called tomosynthesis [30]. Tomosynthesis is entering the clinic as a proposed replacement for mammography [59, 21] and chest radiography [71], where it offers the ability provide improved diagnostic sensitivity and specificity by approximately separating tissues that would be superposed in a single projection image at a lower dose than CT. I have avoided this term in favor of *limited angle tomography* which implies a superset of tomosynthesis that includes other limited angle problems such as 4D CBCT and does not carry any historical baggage of analog tomosynthesis [30], where analog tomosynthesis is an imaging modality where an X-ray source and film cassette are moved to selectively sharpen a single plane in the 3D object while blurring others.

Figures 2.10 and 2.11 demonstrate the effects of limited angular sampling on a digital phantom derived from a clinical CT. The images were reconstructed with 10 iterations of the simultaneous algebraic reconstruction technique (SART) algorithm [1] and one projection taken per degree of angular coverage. The angular coverage was taken about the AP direction. Image quality decreases gracefully as angular sampling decreases, resulting in decreasing resolution in the AP direction. This is known as *artifact spread*. This loss of resolution is supported by the FST since in acquiring a cone beam projection in the AP an approximate plane in the coronal plane of the frequency domain has been measured. This gives rise to a preferred viewing plane that is orthogonal to the central imaging direction. The coronal plane remains more useful for human viewing since the frequency components that give rise to it are better sampled. In current implementations, the imaging directions are usually chosen such that the preferred viewing plane is either the coronal or sagittal because these are anatomically familiar to interpreters. An axial preferred viewing plane is an infeasible geometry for human subjects, since projections would need to be acquired from the superior-inferior (SI) direction. However, it may actually be desirable to have the preferred viewing plane oblique to these planes, due to either the spread of high contrast objects obscuring low contrast objects of interest or due to the specific deformation space of these objects. The downside of this approach is that even though such images may be better interpretable to computers, they may become much more difficult for humans to interpret due to artifact spread from oblique structures and the desire for humans to interpret medical images in the traditional orthogonal planes due to symmetries in the human body. This problem is more a lack of human familiarity, rather than a methodological weakness, and could be overcome with practice. For the time being, humans make the final decisions, and their ability to judge correctly is paramount.

Figure 2.12 shows the response of a limited angle geometry and reconstruction to a small sphere, illustrating the problems of limited angle geometries in isolation. The

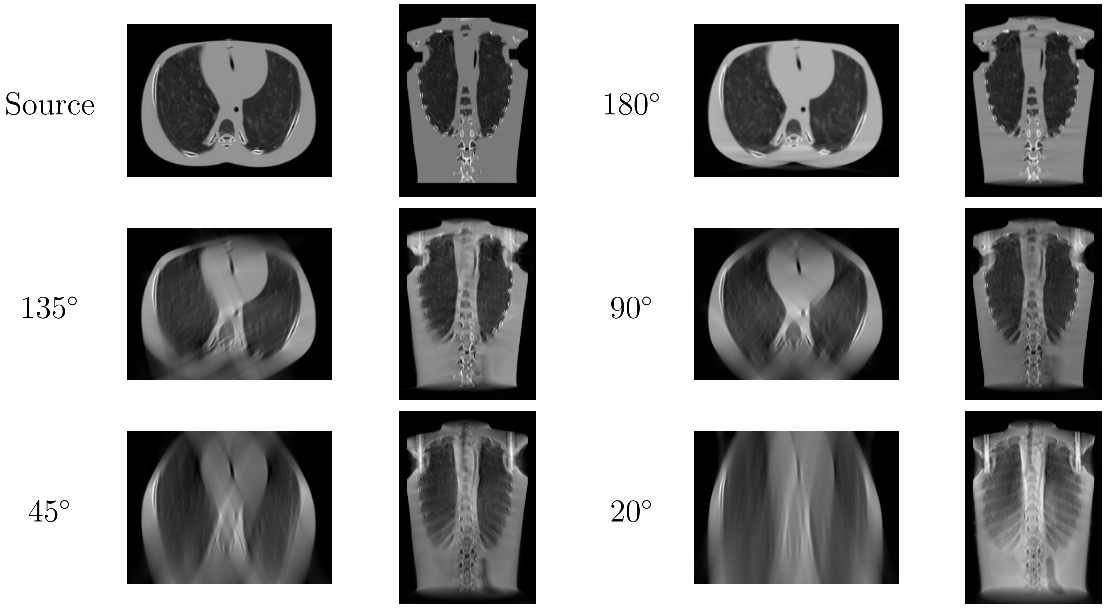


Figure 2.10: Limited angle reconstructions simulated from a high quality clinical FBCT scan of the Rando anthropomorphic torso phantom. Images were reconstructed from projections over the specified angular coverage with one projection per degree of coverage and 10 iterations of SART. Image quality gracefully decreases with increasing artifact spread with a decrease in angular coverage. Even though the axial images with small angular coverage may be difficult to interpret, the coronal images are in the preferred viewing plane and are easily interpretable. Since this is a cone beam acquisition, 180° is not complete angular sampling, analogous to the fan beam case.

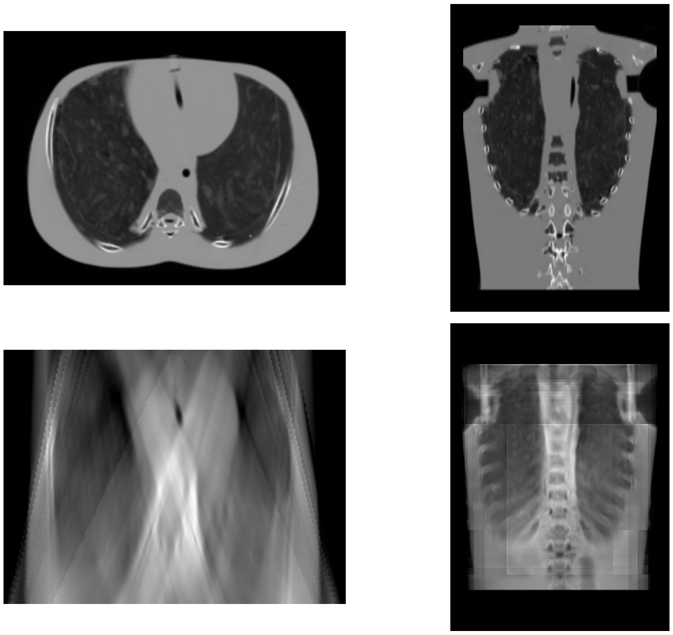


Figure 2.11: Limited angle reconstructions analogous to those in figure 2.10, but simulated with the NST geometry. The line artifacts are the result of truncation where the artifacts indicate the edges of the intersection of the rays from a particular view and the reconstruction domain. The NST geometry is fixed, but angular sampling and image quality could be improved by rotating the linear accelerator gantry.

image is reconstructed from a 20° arc centered about the vertical axis of the image, using the 10 iterations of SART. Artifact spread in the image is obvious. The sphere is spread out in the direction of imaging but remains well localized in the horizontal direction because Fourier space is poorly sampled in the vertical direction but well sampled in the horizontal one. The image also shows a decrease in intensity. This is because the total intensity in the image is conserved under projection and reconstruction, but that intensity is spread out over a larger area. This means that the reconstructed intensities are not comparable to their true values. The intensity that CT, at least approximately, measures is attentuation. Attenuation is a real, quantifiable property of matter interrogated by X-rays. Knowing the calibrated spatial attenuation distribution function of a subject has applications in radiation oncology with application to dose calculation for treatment planning and adaptive therapy and is also important in image registration applications. Because of this phenomena, limited angle images have limited value for dose calculation. It also makes performing registrations between limited angle images and completely sampled images difficult.

2.3 Mathematics of Transformations

This work is concerned with the determination, manipulation, and application of diffeomorphic transformations to medical images. Here, transformations are considered as points in various transformation spaces, from the six dimensional space of rigid transformations to the infinite dimensional space of generalized diffeomorphisms. However, these transformation spaces, other than simple translation, are not Euclidean. They are curved, rather than flat everywhere. Because of this curvature, the familiar mathematical manipulations of points in Euclidean spaces are not applicable to points in non-Euclidean spaces. This is not to say that it is meaningless to treat these transformations as if they were in a Euclidean space, but, in order to manipulate these

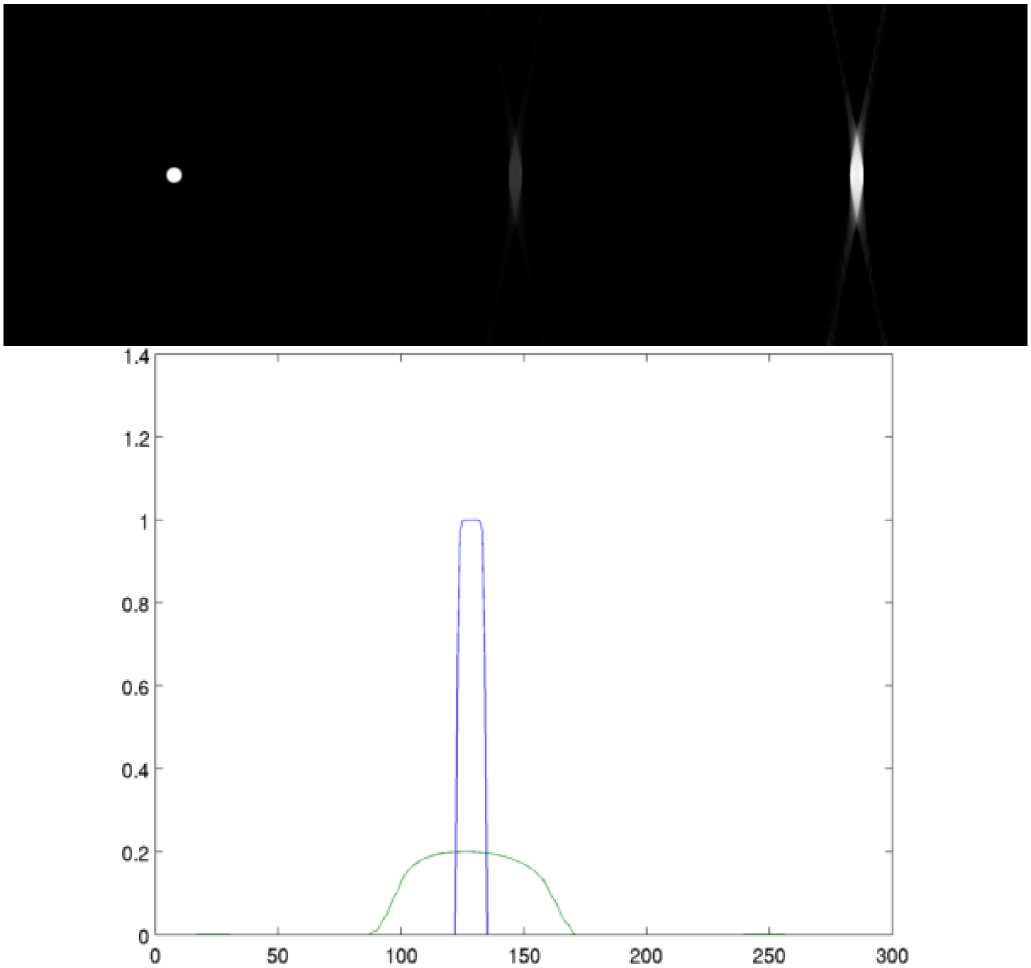


Figure 2.12: Image of a central slice through the sphere response of a sample cone beam geometry with 20° of angular coverage. Left is the source image, center is the response on the same scale as the source image, and right is the contrast enhanced sphere response. Bottom is a vertical profile through the central sphere. The profile through the sphere is shown in blue, and the profile through the sphere reconstructed from the 20° cone beam geometry is shown in green.

transformations properly, the non-Euclidean nature of their spaces must be considered. This section provides the background necessary to extend the reader's understanding of Euclidean spaces to non-Euclidean spaces, namely, by the introduction of Lie groups and the Log-Euclidean framework.

2.3.1 Introduction to Diffeomorphic Transformations

A diffeomorphic transformation is a smooth mapping $\boldsymbol{\phi}$ between two sets of points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ such that each element of $\boldsymbol{x}$ smoothly maps to one and only one element in $\boldsymbol{y}$ and the inverse $\boldsymbol{\phi}^{-1}$ exists and is also smooth. Intuitively, diffeomorphic transformations are those that can be applied to an elastic sheet or putty without folding, tearing, or joining. This is an approximation to the transformations that a human patient can undergo. A diffeomorphism, as it is used here, is simply a subset of all functions from $\mathbb{R}^3$ to $\mathbb{R}^3$. Many parameterizations are possible; the only challenge is ensuring that the function remains smooth and invertible. The parameterization used here is simply a lookup table that is implemented as an ordered set of vectors with tails at $\boldsymbol{x}$ and heads at $\boldsymbol{\phi}(\boldsymbol{x})$. Thus,

$$\{[x_0, y_0, z_0] \rightarrow [x'_0, y'_0, z'_0], \ldots, [x_n, y_n, z_n] \rightarrow [x'_n, y'_n, z'_n]\} \tag{2.31}$$

with some interpolation (usually, bi- or tri-linear) that specifies the vectors intermediate to the samples. The tails are usually specified on an ordered grid whose positions can be inferred from their location within the array. To compose two diffemorphisms $\boldsymbol{\chi}(\boldsymbol{\phi}(\boldsymbol{x}))$ (equivalently, $\boldsymbol{\chi} \circ \boldsymbol{\phi}$) one simply queries the function $\boldsymbol{\phi}$ and then the function $\boldsymbol{\chi}$. This composition is also a diffeomorphism due to the group properties defined below. The second common operation is deformation or transformation of an image. Given an image $I(\boldsymbol{y})$, the image can be brought to coordinate system $\boldsymbol{x}$ with $I(\boldsymbol{\phi}(\boldsymbol{x}))$. That is, iterate over all voxels in $\boldsymbol{\phi}(\boldsymbol{x})$, returning the head of the associated vector, interpolate the intensity of the image I at the corresponding point, and finally store in the correct location in the output array. While not all vector fields represent diffeomorphisms, all diffeomorphisms can be represented by vector fields.

In understanding diffeomorphisms, the Taylor series is a useful tool. Recall that the Taylor series is

$$\boldsymbol{\phi}\left(\boldsymbol{x}\right) = \boldsymbol{\phi}\left(\boldsymbol{x}_0\right) + \operatorname{Jac}\left[\boldsymbol{\phi}\left(\boldsymbol{x}_0\right)\right]\left(\boldsymbol{x} - \boldsymbol{x}_0\right) + \ldots \tag{2.32}$$

plus higher order terms and in the neighborhood of $\boldsymbol{x}_0$. Jac is the Jacobian which is a matrix of derivatives of a function. It is analogous to the gradient of a scalar function. For the 3D diffeomorphisms of interest here the Jacobian matrix is

$$\operatorname{Jac}\boldsymbol{\phi} = \begin{bmatrix} \frac{\partial}{\partial x}\phi_x & \frac{\partial}{\partial y}\phi_x & \frac{\partial}{\partial z}\phi_x \\ \frac{\partial}{\partial x}\phi_y & \frac{\partial}{\partial y}\phi_y & \frac{\partial}{\partial z}\phi_y \\ \frac{\partial}{\partial x}\phi_z & \frac{\partial}{\partial y}\phi_z & \frac{\partial}{\partial z}\phi_z \end{bmatrix} \tag{2.33}$$

and is often computed by finite differences. Much intuitive information about the transformation can be learned from the Jacobian matrix. In the case where all higher order terms of the Taylor series are $\mathbf{0}$, the transformation is linear and has the form $\boldsymbol{\phi}\left(\boldsymbol{x}\right) = \boldsymbol{M}\boldsymbol{x} + \boldsymbol{t}$, where $\boldsymbol{M}$ is a matrix and $\boldsymbol{t}$ is a translation vector. If higher order terms are not $\mathbf{0}$, the constant term and Jacobian matrix form a locally linear approximation to the transformation.

The determinant of $\boldsymbol{M}$ at $\boldsymbol{x}$, $\det \boldsymbol{M}$ defines the scaling of space at $\boldsymbol{x}$. If $\det \boldsymbol{M} = \frac{1}{2}$, a volume element there will be shrunk to half its volume. When $\det \boldsymbol{M} = 0$, the volume element has shrunk to zero. Intuitively, such an element would not be invertible. This agrees with the knowledge that a matrix with determinant 0 is not invertible. A negative determinant indicates a “flip”, for example from a right-handed to a left-handed coordinate system. A transformation with both positive and negative Jacobian determinants is not a diffeomorphism since this would indicate folding somewhere.

The transformation that is most common in studies within the same subject is the rigid transformation. In this case, $\boldsymbol{M}$ is orthonormal and constant over space. It provides a translation of the origin and rotation of the basis vectors (coordinate axes) maintaining their orthogonality. This transformation in 3D has 3 degrees of freedom for translation and 3 for rotation. Such a transformation is typically estimated first when aligning a previous image of a subject with a new image of a subject. In inter-subject studies, additional degrees of freedom may be added. These include rigid plus global scaling, rigid plus per coordinate scaling, and fully affine (any matrix with a positive determinant).

Affine transformations with translations are often represented with a single matrix using homogeneous coordinates. A vector in 3D homogeneous coordinates is represented $\boldsymbol{x} = [x_0, x_1, x_2, 1]^T$. The vector is then operated upon by a 4×4 matrix with the following form:

$$\boldsymbol{y} = \boldsymbol{M}'\boldsymbol{x} \tag{2.34}$$

$$\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{M} & \boldsymbol{t} \\ \boldsymbol{0} & 1 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ 1 \end{bmatrix} \tag{2.35}$$

where $\boldsymbol{M}$ is a 3×3 matrix representing an affine transform and $\boldsymbol{t}$ is a 3×1 vector representing a translation following the rotation. In this dissertation, the final row is typically $[0, 0, 0, 1]^T$, but it can be used to represent a perspective transformation (where parallel lines do not remain parallel, as is the case for all affine transformations).

Such transformations find many applications but are not commonly used as volume transformations in medical image analysis.

2.3.2 Lie Groups and Manifolds

Non-Euclidean Spaces and Tangent Planes

In elementary mathematics, spaces are often assumed to be Euclidean, that is, flat and spanned globally by a vector space, $\mathbb{R}^n$. This work is concerned with transformations which exist in non-Euclidean spaces. Many such non-Euclidean structures exist, and they cannot necessarily be properly analyzed when their non-Euclidean nature is not considered. As a first example, consider the set of points on a sphere. A sphere is a 2D *manifold* $\mathbb{S}^2$, because any point on it can be specified by 2 coordinates, embedded in $\mathbb{R}^3$. A manifold is a set of points embedded in a space where, in the neighborhood of each point, the point set appears flat. That is, a tangent plane can be placed on it. In a Euclidean space, that tangent plane is the Euclidean space itself. A sphere is not flat everywhere, but a map (in the sense of cartography and the more general mathematical sense) can be constructed at each point. The sphere is therefore a non-Euclidean manifold.

The sphere demonstrates both the problems with treating non-Euclidean spaces as if they were Euclidean and possible solutions to these problems. If one wishes to find the population center of Asia using the standard Euclidean rules, the curvature of the Earth means that the result will be somewhere deep underground, rather than on the surface, which is where the more correct solution exists. However, on the scale of a single town, this approximation is appropriate. This ability to be locally approximated by a Euclidean tangent space is the differentiability of the manifold. The key idea is that because the manifold is differentiable, the neighborhood around each point can be mapped to a tangent Euclidean space where analysis can be performed. All the non-Euclidean manifolds to be discussed here are differentiable everywhere. They have no holes, corners, or edges. Furthermore, they are also *Riemannian* because a notion of distance can be defined on the manifold. A curve on the surface of the manifolds discussed in this work has a well defined length. There is a distance between two rotations or diffeomorphisms in the same sense that there is a distance between two cities. There is, of course, more than one distance between two points on a manifold, dependent on the path taken, but the path that gives the shortest distance while remaining in the manifold has special meaning. This shortest path is called the geodesic. Geodesics are the notion of a straight line generalized to a curved space. In Euclidean space, the geodesic is the straight line. On a sphere, the geodesics are the great circles (i.e., shown by lines of longitude but not lines of latitude, excepting the equator).

In order to extend Euclidean tools to non-Euclidean space, one can take advantage of Riemannian structure of the manifold. As described above, the key is to develop a mapping between each point on the manifold and a tangent Euclidean space at that point. With proper modification, Euclidean operations can be performed in this tangent space. Operations appropriately performed in this tangent space produce results that both remain in the space and take this manifold (equivalently, geodesic or Riemannian) distance into account. The tangent space at a point is the set of directions in which an infinitesimal step can be taken while still remaining on the manifold and has the same dimensionality as the manifold. Many possible mappings from the manifold to the tangent space exist, but the particular mapping of interest is the one that maps nearby points to the tangent such that the manifold distance from the tangent point to the point of interest becomes the Euclidean distance in the tangent space. This mapping from the manifold to a tangent Euclidean space is known in the literature as the *Logarithm* and has an inverse known as the *Exponential*. These are generalizations of the standard exponential and logarithm and are distinguished by the use of capital letters and $\mathfrak{log}$ rather than log. The similarities between the specific case and the generalization are described below.

Logarithms and Exponentials

The Log and Exp maps are most easily visualized with a descriptive example of the earth, shown in figure 2.13. The great circle distance from Chapel Hill, NC, to any point on the Earth is the Euclidean distance in the tangent space. Furthermore, the path from Chapel Hill to any point on the Earth [1] is a straight line in the tangent space and maps to a great circle using the Exp mapping. However, the Log and Exp mappings are unique to each point, so a path not containing Chapel Hill only maps to a geodesic in special cases. The great circle route from New York to San Francisco is not a straight line in the tangent space at Chapel Hill.

As noted above, the Exp and Log mappings are generalizations of the familiar scalar exponential and logarithm. Specifically, the familiar log maps the positive reals $\mathbb{R}^+$ with distance metric $d(a, b) = |\log a - \log b|$ to the reals $\mathbb{R}$ with a standard Euclidean distance. This space, $\mathbb{R}^+$ with $d(a, b) = |\log a - \log b|$, is an appropriate one in which to consider transformations the consist only of uniform scaling with a positive scalar. Further similarities between the standard log/exp and Log/Exp will be explored as additional concepts are introduced.

Groups

The mathematics here are used to manipulate transformations, and the two fundamental operations for diffeomorphic transformations are composition and inversion. Consider the composition of two diffeomorphic uniform scaling transformations, $s_3 = s_1 s_2$.

[1] Technically, the Log mapping does not exist at the antipode, but the Earth is technically an oblate spheroid, for which a similar Log mapping exists.

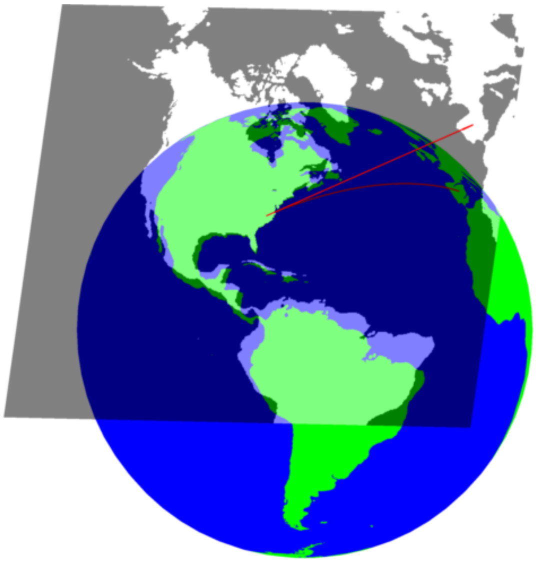


Figure 2.13: The tangent space is located at Chapel Hill, North Carolina. Using the Logarithmic mapping, points are mapped from the Earth's surface to the tangent plane such that Euclidean distances correspond to geodesic distances. A vector is drawn in the tangent plane from Chapel Hill, NC to Madrid, Spain which maps the vector to an equal length geodesic on the Earth's surface.

This is a binary operation, one taking two elements and producing a third. There is also an operation, multiplication, to compose the two elements. It is obvious that for any $s_1, s_2 \in \mathbb{R}^+$ composed under multiplication s_3 must also be in $\mathbb{R}^+$. Similarly, the inverse of a uniform scaling transformation $(s^{-1}) = \frac{1}{s}$ is also in $\mathbb{R}^+$. These properties are only preserved when the composition operator is multiplication and the inverse operator is division. If the operators were addition and subtraction, respectively, the result of manipulations will not always be in $\mathbb{R}^+$. This is because $\mathbb{R}^+$ forms a *group* under multiplication with inverse division. Additionally, the distance metric $d\left(a,b\right) = \left|\log a - \log b\right|$ is a sensible choice for uniform scaling since a scaling of 2 is intuitively the same distance from the identity, a scaling of 1, as is a scaling of $\frac{1}{2}$.

A group is a set combined with a binary operation, analogous to multiplication, such that any two elements a and b combined using the binary operation produces another element that is also in the group. This property is called *closure*. Groups are also associative $a \cdot (b \cdot c) = (a \cdot b) \cdot c$, have a unique identity element $a = Id \cdot a = a \cdot Id$, and have a unique inverse $Id = a \cdot a^{-1} = a^{-1} \cdot a$. The $\mathbb{R}^+$ example in the previous paragraph has all these properties. Some groups considered in this work are the special orthogonal group $\mathbb{SO}^3$ (rotations in 3D), the special Euclidean group $\mathbb{SE}^3$ (rotations and translations in 3D), and the general linear group $\mathbb{GL}^3$ (affine transformations in 3D). The group operator for these matrix groups,when represented by matrices is matrix multiplication, and the inverse from matrix inversion exists for all elements in each of the groups. Furthermore, these are all nested sub-groups of diffeomorphisms $\mathbb{DIFF}^3$ (infinite dimensional non-rigid transformations in 3D) in that all elements in $\mathbb{DIFF}^3$ contains all elements of $\mathbb{GL}^3$, which contains all elements of $\mathbb{SE}^3$, which contains all elements of $\mathbb{SO}^3$. However, the group operation for $\mathbb{DIFF}^3$ is functional composition and inversion, which reduces to matrix multiplication and inversion for the $\mathbb{GL}^3$, $\mathbb{SE}^3$, and $\mathbb{SO}^3$. Finally, the transformations discussed in this work are also Lie groups[2] because they satisfy the properties of a group in general and are also Riemannian manifolds.

To summarize with rotations $\mathbb{SO}^3$ as an example, given any two rotations, they can be combined using matrix multiplication to obtain a new transformation that is also a rotation; there is a unique identity element (no rotation, the identity matrix); and for every rotation, there is another rotation that when combined with that rotation results in the identity rotation. There also exist Exp and Log maps for elements in the group. At a given point at, for example, the identity, a tangent plane can be constructed that maps from any rotation to a Euclidean space. For rotations and a Log mapping at the identity, this tangent space is the axis-angle representation of the rotation (an angle of rotation multiplied by a unit vector axis of rotation). In the tangent axis-angle

[2] Technically, Lie groups are finite dimensional because certain theorems about them are only true in the finite dimensional case. However, diffeomorphisms resemble Lie groups sufficiently that they can be considered as such for the purposes of this work. They can be distinguished with the term *infinite dimensional Lie groups*.

space, any vector in $\mathbb{R}^3$ maps to a rotation. Because this space is Euclidean, rather than the non-Euclidean space of orthonormal matrices, Euclidean operations, such as averaging can be performed, that respect the inherent distance metric in $\mathbb{SO}^3$ and that are guaranteed to produce another element in $\mathbb{SO}^3$ as the result. The alternative approach to averaging rotations would be to simply average orthonormal matrices. This is not guaranteed to produce an orthonormal matrix and, in general does not do what it is intended to (as demonstrated in chapter 4).

Definitions of the Exp and Log Mappings

As mentioned above, the group Exponential and Logarithm are generalizations of the standard exponential, and they share similar definitions. Since the Exponential mapping is more commonly applied to map from a Euclidean space into the desired group, it is defined here with the understanding that the Logarithm is its inverse. Consider the Taylor series definition of the exponential of a scalar $\boldsymbol{M}$

$$\mathfrak{exp}\,\boldsymbol{M} = \sum_{i=0}^{\infty} \frac{1}{i!}\boldsymbol{M}^i \tag{2.36}$$

If $\boldsymbol{M}$ is in $\mathbb{R}$, $\boldsymbol{M}^i$ signifies repeated multiplication, and $\boldsymbol{M}^0$ is the identity for multiplication 1, then the result is the standard exponential which forms the group of uniform scaling described above. Now let us generalize $\boldsymbol{M}$ to a matrix, if $\boldsymbol{M}$ is a skew-symmetric matrix which is one of the form

$$\boldsymbol{a} = \begin{bmatrix} a_x & a_y & a_z \end{bmatrix} \tag{2.37}$$

$$\boldsymbol{M} = \begin{bmatrix} 0 & -a_z & a_y \\ a_z & 0 & -a_x \\ -a_y & a_x & 0 \end{bmatrix} \tag{2.38}$$

where $\boldsymbol{a}$ is the axis-angle representation of rotation, $\boldsymbol{M}^i$ signifies repeated matrix multiplication, and $\boldsymbol{M}^0$ is the identity for matrix multiplication (the identity matrix), $\mathfrak{exp}\,\boldsymbol{M}$ is an orthonormal matrix, a rotation. By changing the form of the matrix, the different matrix group Exponentials can be defined. For example, for a general matrix $\boldsymbol{M}$, $\mathfrak{exp}\,\boldsymbol{M}$ is in $\mathbb{GL}$. If the matrix is of the form

$$\boldsymbol{a} = \begin{bmatrix} a_x & a_y & a_z & v_x & v_y & v_z \end{bmatrix} \tag{2.39}$$

$$\boldsymbol{M} = \begin{bmatrix} 0 & -a_z & a_y & v_x \\ a_z & 0 & -a_x & v_y \\ -a_y & a_x & 0 & v_z \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{2.40}$$

with axis-angle representation and velocity vector (more on velocity below), the result is in $\mathbb{SE}^3$, rigid transformations. This is the group that is used in poly-rigid transformations, and the matrix exponential and logarithm have convenient closed form solutions for $\mathbb{SE}^3$ [52]. To form $\mathbb{DIFF}^3$, $\boldsymbol{M}$ is a 3D velocity vector field, $\boldsymbol{M}^i$ is repeated functional composition, and $\boldsymbol{M}^0$ is the identity for functional composition.

The Exponential mapping for $\mathbb{DIFF}^3$ is also defined as the solution to an ordinary differential equation (ODE) of form

$$\dot{\boldsymbol{x}} = \boldsymbol{v}\left(\boldsymbol{x}\right) \tag{2.41}$$

$$\boldsymbol{\phi}\left(\boldsymbol{x}, t\right) = \mathfrak{exp}\, t\boldsymbol{v}\left(\boldsymbol{x}\right) \tag{2.42}$$

for the case where $\boldsymbol{v}$ is stationary (constant in time) and where the initial conditions are typically $\boldsymbol{x}$, the domain of the problem. The desired solution usually occurs at $t = 1$, but varying t moves along the geodesic between the identity $\boldsymbol{x}$ and $\boldsymbol{\phi}\left(\boldsymbol{x}, 1\right)$ (and beyond as t further increases). This is known as the flow of a one-parameter subgroup.

Similarly, by restricting the form of the velocity vector field (VVF), the exponentials for the sub-groups of $\mathbb{DIFF}^3$ are obtained.

$$\dot{\boldsymbol{x}} = \boldsymbol{M}\boldsymbol{x} \tag{2.43}$$

$$\boldsymbol{\phi}\left(\boldsymbol{x}, t\right) = \mathfrak{exp}\ \left(t\boldsymbol{M}\right)\boldsymbol{x} \tag{2.44}$$

Strictly speaking, $\boldsymbol{\phi}$ is the transformation as a function of $\mathbb{R}^3 \rightarrow \mathbb{R}^3$ but is equivalent to and more succinctly expressed as simply the matrix $\mathfrak{exp}\,\boldsymbol{M}$. For the scalar case, the result is the standard exponential.

While the Exponential and Logarithm have closed form solutions for $\mathbb{SE}^3$, the Exponential of a smooth function $\boldsymbol{v}\left(\boldsymbol{x}\right)$ does not have a closed form solution in general. As such it must be computed numerically. In order to compute the Exponential, any of the myriad of numeric integration methods, or even the Taylor series definition of the Exponential seen in equation (2.36), can be used. In [2], a fast *scaling and squaring* method was proposed that generalizes the popular method for computing the general matrix exponential to non-linear transformations. This is a fast variant of the first order Euler's method for 2^N steps. First, the transformation is scaled $\frac{\boldsymbol{v}(\boldsymbol{x})}{2^N}$, so that it is close to $\mathbf{0}$. When the transformation is small, the first order Taylor series approximation to the Exponential $\boldsymbol{Id} + \frac{\boldsymbol{v}(\boldsymbol{x})}{2^N}$ is sufficiently accurate. The transformation is then recursively squared (self composed) n times to compute the exponential of $\boldsymbol{v}\left(\boldsymbol{x}\right)$. Since any sufficiently small transformation is diffeomorphic, $\frac{\boldsymbol{v}(\boldsymbol{x})}{2^N}$ is diffeomorphic for sufficiently large N. Then, because diffeomorphisms form a group under composition, $\mathfrak{exp}\,\boldsymbol{v}\left(\boldsymbol{x}\right)$ computed by this recursive self composition method is also diffeomorphic.

The issue with this method is that this work was implemented for parallel GPUs using Compute Unified Device Architecture (CUDA) (see appendix A). These devices have hardware accelerated linear interpolation, which is needed for functional composition of vector fields as implemented in this work. However, the hardware implementation in these devices is limited in precision. The scaling and squaring method requires time steps on the order of $\frac{1}{2^7}$ or $\frac{1}{2^8}$ in order to produce good results, but these time steps are too small to be accurate respecting the limited precision with which linear interpolation is implemented in GPU hardware. In order to take advantage of the hardware interpolation, a larger time step should be used, which means that a higher order numerical integration method should be used. Higher order numerical integration methods have less error for a given step size than lower order numerical integration methods. In this work, I have used the popular 4th order Runge-Kutta method, which has total error on the order of h^4 as opposed to h^1 provided by Euler's method, where h is the step size. Using the Runge-Kutta method means that accurate exponentials can be calculated using only 32 steps.

The 4th order Runge-Kutta method solves the initial value problem in equation 2.41 for each point in the domain independently, where the initial value is the identity. For N steps with step size h such that $hN = 1$, the Runge-Kutta method for computing Exponentials of diffeomorphisms is

$$\boldsymbol{y}_{n+1} = \boldsymbol{y}_n + \frac{1}{6}h\left(\boldsymbol{k}_1 + 2\boldsymbol{k}_2 + 2\boldsymbol{k}_3 + \boldsymbol{k}_4\right) \tag{2.45}$$

$$t_{n+1} = t_n + h \tag{2.46}$$

where

$$\boldsymbol{k}_1 = \boldsymbol{v}\left(\boldsymbol{y}_n\right)) \tag{2.47}$$

$$\boldsymbol{k}_2 = \boldsymbol{v}\left(\boldsymbol{y}_n + \frac{h}{2}\boldsymbol{k}_1\right) \tag{2.48}$$

$$\boldsymbol{k}_3 = \boldsymbol{v}\left(\boldsymbol{y}_n + \frac{h}{2}\boldsymbol{k}_2\right) \tag{2.49}$$

$$\boldsymbol{k}_4 = \boldsymbol{v}\left(\boldsymbol{y}_n + h\boldsymbol{k}_3\right) \tag{2.50}$$

The temporal aspect of the approach has been left out of equation (2.47) because $\boldsymbol{v}\left(\boldsymbol{x}\right)$ is constant in time. This method is sufficiently fast on the GPU and is trivially implemented.

The previous paragraphs demonstrated calculation of Exponentials of general VVFs, mapping them to diffeomorphisms. This can be done quickly and accurately. However, calculation of Logarithms of diffeomorphisms was not demonstrated. In order to compute Logarithms, an optimization based approach can be used along with methods described in [10]. However, computing Logarithms of diffeomorphisms is undesirable due to its computational cost, and it will be shown that taking Logarithms of diffeomorphisms is not necessary for this work.

This paragraph introduces two additional facts about the framework. The first is the inverse property. The Exponential of a negative Logarithm is the inverse of the Exponential $\mathfrak{exp}\left(-\boldsymbol{a}\right) = \left(\mathfrak{exp}\,\boldsymbol{a}\right)^{-1}$. This is true for all groups and all $\boldsymbol{a}$. It can be observed from the uniform scaling group with inverse division by $\exp\left(-a\right) = \frac{1}{\exp a} = \frac{1}{b}$, where b is the standard domain representation equivalent to a. This becomes useful for diffeomorphisms because inversion in the Log domain followed by Exponentiation is extremely inexpensive when compared to inverting diffeomorphisms in the standard domain. Rather than simplifying, the second fact complicates. The familiar identity from the standard exponential $\mathfrak{log}\left[\mathfrak{exp}\,\boldsymbol{a}\ \cdot \mathfrak{exp}\,\boldsymbol{b}\right] = \boldsymbol{a} + \boldsymbol{b}$ does not hold unless $\mathfrak{exp}\,\boldsymbol{a}\,\cdot$

$\mathfrak{exp}\,\boldsymbol{b} = \mathfrak{exp}\,\boldsymbol{b} \cdot \mathfrak{exp}\,\boldsymbol{a}$, the group is commutative (*Abelian*). The uniform scaling group is commutative. Matrix groups and diffeomorphisms are not unless $\boldsymbol{a}$ and $\boldsymbol{b}$ are members of the same one-parameter subgroup (lie along the same geodesic).

Computing the Mean of Samples in a Non-Euclidean Space

Using the tools developed above, a analogous notion of average can be developed in the Non-Euclidean space. In non-Euclidean spaces, the generalization of the arithmetic mean is called the Fréchet mean. The methods used here are normally referred to as the Log-Euclidean framework, since the distances used are Euclidean distances between Logarithms, and can be used to extend other techniques developed for Euclidean spaces. The Fréchet mean is defined in the same way as a Euclidean mean

$$\overline{\boldsymbol{b}} = \arg\min_{\hat{\boldsymbol{b}} \in \mathbb{G}} \sum_i d\left(\boldsymbol{b}_i, \hat{\boldsymbol{b}}\right)^2 \tag{2.51}$$

for group $\mathbb{G}$ and distance metric d. That is, the element in the space that minimizes the sum of squared geodesic distances between that element and all the other elements in the sample. This problem can be rewritten using notation from the Log-Euclidean framework. Consider the $\boldsymbol{b}_i$'s by their Log domain representation $\mathfrak{log}\,\boldsymbol{b}_i = a_i$, mapping $\mathbb{G} \to \mathbb{R}^n$. This enables standard Euclidean techniques to be used in solving (2.51).

$$\overline{\boldsymbol{a}} = \arg\min_{\hat{\boldsymbol{a}}} \sum_i \left|\left|\mathfrak{log}\,\left[\left(\mathfrak{exp}\,\boldsymbol{a}_i\right) \cdot \left(\mathfrak{exp}\,\hat{\boldsymbol{a}}^{-1}\right)\right]\right|\right|^2 \tag{2.52}$$

The desired solution is then $\overline{\boldsymbol{b}} = \mathfrak{exp}\,\overline{\boldsymbol{a}}$. As a concrete example, consider first the $\mathbb{R}^+$ space with group operator multiplication, inverse division, and the standard exponential and logarithm. Substituting the inverse, Exponential, and Logarithm (2.52),

$$\begin{aligned} \overline{a} &= \arg\min_{\hat{a}} \sum_i \left(\log \frac{\exp a_i}{\exp \hat{a}} \right)^2 & (2.53) \\ &= \arg\min_{\hat{a}} \sum_i (a_i - \hat{a})^2 & (2.54) \end{aligned}$$

Finding the optimum, using the standard technique

$$\begin{aligned} 0 &= \frac{\partial}{\partial \hat{a}} \sum_i (a_i - \hat{a})^2 & (2.55) \\ &= -2 \sum_i (a_i - \hat{a}) & (2.56) \\ \hat{a} &= \frac{1}{N} \sum_i a_i & (2.57) \end{aligned}$$

Finally, mapping back into the standard domain

$$\overline{b} = \exp \sum_i \log b_i \qquad (2.58)$$

Equation (2.58) calculates the mean of the uniform scaling group, acknowledging its distance function and always producing results that are also in the group. By substituting the appropriate Exp and Log mappings, (2.52) can then be used to find an iterative process for determining the mean of rigid transformations and diffeomorphisms.

$$\overline{\boldsymbol{a}}_{k+1} = \frac{1}{N} \sum_i \mathfrak{log} \left[(\mathfrak{exp}\, \boldsymbol{a}_i) \cdot (\mathfrak{exp}\ (-\overline{\boldsymbol{a}}_k)) \right] \cdot (\mathfrak{exp}\, \overline{\boldsymbol{a}}_k) \qquad (2.59)$$

Starting with an initial estimate of the mean, pre-apply the inverse of this estimate and take the Logarithm. Each of the samples is in a tangent Euclidean space at the identity. Find their mean in the Log domain, and use this to update the estimate of the mean. This algorithm converges quickly for practical cases.

However, equation (2.59) requires computation of Logarithms. For elements in $\mathbb{SE}^3$, the actual formulas are simple enough that the exact $\mathfrak{log}\left[\mathfrak{exp}\,\boldsymbol{a}\cdot\mathfrak{exp}\,\boldsymbol{b}\right]$ method can be used. However, for diffeomorphisms, Logarithms are computationally costly operations. In order to avoid computing Logarithms of diffeomorphisms, the Baker-Campbell-Hausdorff formula (BCH) can be used to approximate $\mathfrak{log}\left[\mathfrak{exp}\,\boldsymbol{a}\ \cdot\mathfrak{exp}\,\boldsymbol{b}\right]$. Using the BCH to compute Fréchet means of diffeomorphisms is described in [9].

The BCH is an infinite sum that computes $\mathfrak{log}\left[\mathfrak{exp}\,\boldsymbol{a}\cdot\mathfrak{exp}\,\boldsymbol{b}\right]$. The BCH relies on the bi-linear Lie bracket or commutator, which describes the extent to which the group operation fails to be commutative. It is defined as

$$[\boldsymbol{a},\boldsymbol{b}] = \left[\mathrm{Jac}\,\boldsymbol{b}\right]^T\boldsymbol{a} - \left[\mathrm{Jac}\,\boldsymbol{a}\right]^T\boldsymbol{b} \tag{2.60}$$

where the partial derivatives required for the Jacobian matrix are calculated using the finite difference approximations. In the center of the volume, central differences are used and at the edges, the appropriate forward or backwards difference is used. If the group is commutative, the Lie bracket is zero, and the BCH reduces to the familiar $\mathfrak{log}\left[\mathfrak{exp}\,\boldsymbol{a}\cdot\mathfrak{exp}\,\boldsymbol{b}\right] = \boldsymbol{a}+\boldsymbol{b}$ The first few terms of the BCH are

$$\mathfrak{log}\left[\mathfrak{exp}\,\boldsymbol{a}\,\mathfrak{exp}\,\boldsymbol{b}\right] = \boldsymbol{a}+\boldsymbol{b}+\frac{1}{2}\left[\boldsymbol{a},\boldsymbol{b}\right]+\frac{1}{12}\left[\boldsymbol{a},\left[\boldsymbol{a},\boldsymbol{b}\right]\right]-\frac{1}{12}\left[\boldsymbol{b},\left[\boldsymbol{a},\boldsymbol{b}\right]\right]+\ldots \tag{2.61}$$

with higher terms including increasing nesting of Lie brackets. Since the BCH is a differential operator and is nested, it requires knowledge of high order derivatives. These higher order derivatives are expected to be small, but there are errors associated with finite difference approximations and their sensitivity to noise. Firstly, the series must be truncated to some number of terms (taken to be the first three in this work), and, secondly, the accuracy to which this equation can be calculated is limited by the energy of the transformation for numerical stability reasons. Equation (2.59) can be rewritten with the BCH approximation to the Logarithm for better performance.

2.4 Registration

Registration is the process of estimating the parameters of a geometric transformation such that a transformed image best matches some reference image. Registration places both images into the same coordinate system so that points in the reference and transformed image correspond with each other. There are three main parts to a registration algorithm. The first is the family from which the transformation is to be selected. These can be rigid, similarity, affine, various parameterizations of non-rigid, and others. Typically, rigid or affine registration is performed before any non-rigid registration. Rigid and affine registration is a common problem that has been discussed thoroughly in the literature; it is not discussed here. For the purposes of this section, the transformation is non-rigid and represented by a vector field, rather than a B-spline [62] or other representation. Further details about the representation will be explained in section 2.4.1. After selecting the transformation, the second choice is that of the image distance. This is determined by the requirements of the modality but can be simple to change. Frequently, for intra-modality registration (e.g., CT to CT rather than CT to MR), sum of squared differences (SSD), or least squares, is sufficient. The third is regularization. Regularization prevents arbitrary rearrangement of voxels by suggesting that a better transformation is smoother in some sense. Regularization is not typically necessary for low dimensionality transformations such as affine or rigid due to the few degrees of freedom in such transformations. With these transformations, infeasible transformations are not very accessible and are not likely to be suggested by the image data. On the other hand, non-rigid registrations have very many degrees of freedom and most transformations in the space of non-rigid transformations are infeasible. Furthermore, without regularization in non-rigid registration, there is typically not a unique *best* transformation.

2.4.1 Non-rigid Registration

This section discusses diffeomorphic image registration methods. The problem of non-rigid registration is the problem of finding a credible transformation between two images, and, in many cases, a credible transformation is diffeomorphic. These diffeomorphic image registration methods ensure that the resulting transformation is diffeomorphic by construction. Many successful and popular methods do not share this property. There, invertibility is suggested only by smoothing or parameterization choices, or differentiability is not desired, as in the case of cutting or sliding. However, since invertibility and smoothness are requirements for this work, it is desirable to have the additional mathematical rigor from a truly diffeomorphic method. The three registration methods that are discussed here are diffeomorphic demons [69], symmetric Log demons [70], and large deformation diffeomorphic metric mapping (LDDMM) [17]. Symmetric Log demons is an extension to Log demons, enabled by the Log-Euclidean framework. Without this symmetric extension, Log demons is explained along with diffeomorphic demons and LDDMM. The symmetric extension is described separately.

Non-rigid registration problems are typically stated

$$\hat{\boldsymbol{\phi}} = \arg\min_{\boldsymbol{\phi}} d\left(I\left(\boldsymbol{x}\right), J\left(\boldsymbol{\phi} \circ \boldsymbol{x}\right)\right) + R\left(\boldsymbol{\phi}\right) \tag{2.62}$$

where d is the image distance term, R is the regularization term, and $\boldsymbol{\phi}$ maps J to I.

Image Distance

The image distance term is a function of two images in the same space: the fixed template image I and the moving image to be registered J, to which the transformation $\boldsymbol{\phi}$ is applied. It is a scalar function of how well all voxels in the two images correspond to each other. SSD is typically a effective first choice for CT to CT registrations.

$$d\left(I\left(\boldsymbol{x}\right), J\left(\boldsymbol{\phi} \circ \boldsymbol{x}\right)\right) = \int_{\Omega} \left|\left|I\left(\boldsymbol{x}'\right) - J\left(\boldsymbol{\phi} \circ \boldsymbol{x}'\right)\right|\right|^{2} \; d\boldsymbol{x}' \tag{2.63}$$

However, SSD is typically only successful when corresponding positions have equivalent intensities, excluding Gaussian noise. In MR or in the case of incorrectly calibrated of CTs, there can be a linear scaling $I'\left(\boldsymbol{x}\right) = aI\left(\boldsymbol{x}\right) + b$. Normalized cross correlation (NCC) overcomes this issue by being insensitive to such linear scaling and shifting. Note that NCC and many other image "distance" terms are not necessarily proper metrics. Both of the image distance functions have the weakness that they are very local. That is, they only depend on the value of the individual voxels being compared (with the exception of NCC which also uses all the voxels to get a global intensity scale and shift). Better image distance terms make use of information from a region about each voxel pair. Feature-based image distance terms construct a vector of linear features for each voxel, typically features like Gaussian derivatives and Gabor features, both taken over a range of scales. These features summarize the neighborhood of each voxel, increasing its distinctiveness. However, feature-based image distance terms suffer from the problem of feature selection (which features are important) and feature scaling (what relative importance, or weight, should be ascribed to each feature). Even better image distance terms may simply consider the entire patch about each voxel, that is, all the voxels in a neighborhood about the voxel under consideration. This is the approach used by sum of Gaussian weighted local normalized cross correlation (GW-LNCC)[11], used in this work.

The problem with image distance in CT to CT registration in the pelvis is low contrast, even though it is often used successfully. In SSD, the cost function is dominated by high contrast differences, such as those between bone and soft-tissue or soft-tissue and air. If voxels being compared have a small intensity difference, they can contribute very little to the image distance term, even if they don't, in fact, correspond well. By considering a patch about each voxel and normalizing for the contrast in that patch, differences in regions of low contrast can be made to have a comparable effect on the image distance term as differences in regions of high contrast.

GW-LNCC is defined as a sum of NCC computed over dense, overlapping, local patches (one patch about every voxel). While this may seem computationally expensive, the method can be computed quickly with convolution. Furthermore, common kernels, including the box and Gaussian, which corresponding to the weighting of NCC, are separable. Using separable kernels decreases the computational complexity in 3D from $O\left(K^3N\right)$ to $O\left(3KN\right)$ for an image with N voxels and a convolution kernel of width K. Box and Gaussian kernels also have fast recursive infinite impulse response (IIR) implementations, which make computation independent of kernel width. This can greatly speed up computation over the standard finite impulse response (FIR) implementation when the kernel is large [20]. GW-LNCC has the additional benefit of being insensitive to slowly varying additive and multiplicative intensity transformations. This can overcome certain types of distortion, such as scatter and MR bias field, and reduce the dominance of high contrast anatomy in the image distance term, as mentioned above.

Computing GW-LNCC requires several convolutions to compute the intermediate images.

$$\overline{I}_\sigma = G_\sigma * I\left(\boldsymbol{x}\right) \tag{2.64}$$

$$\overline{J'}_\sigma = G_\sigma * J\left(\boldsymbol{\phi} \circ \boldsymbol{x}\right) \tag{2.65}$$

$$\langle I, J' \rangle_\sigma = G_\sigma * \left[\left(I\left(\boldsymbol{x}\right) - \overline{I}_\sigma\right)\left(J\left(\boldsymbol{\phi} \circ \boldsymbol{x}\right) - \overline{J'}_\sigma\right)\right] \tag{2.66}$$

$$\mathrm{LNCC}_\sigma\left(I, J'\right) = \frac{\langle I, J' \rangle_\sigma}{\sqrt{\langle I, I \rangle_\sigma \langle J', J' \rangle_\sigma}} \tag{2.67}$$

where $G_\sigma *$ signifies convolution with a Gaussian with parameter σ and $\langle \cdot, \cdot \rangle$ signifies the inner product. Each term returns an image and LNCC is defined at every point with proper boundary conditions. The desired scalar result is the sum or integral over the image. Since GW-LNCC, like NCC, increases with increasing similarity, the actual image distance is usually taken to be the negative of the result to be more similar to the standard SSD.

Regularization

After the image distance term, the regularization method is selected. This is the major component of a registration method and typically gives the algorithm its name. Each of the three methods; diffeomorphic demons, Log demons, and LDDMM; discussed in this section have distinct regularization strategies.

Both diffeomorphic demons and Log demons share some basic similarities. Both regularization terms penalize the gradient magnitude of the representation of the transformation. The main difference between the two is that representation. For diffeomorphic demons, the representation of the transformation is a displacement vector field (DVF). For Log demons, it is a Log domain VVF, which is Exponentiated to get a DVF. That is, R for diffeomorphic demons is

$$R\left(\boldsymbol{\phi}\right) = \frac{1}{\sigma^2}\int_{\Omega} \left|\left|\nabla \boldsymbol{\phi}\left(\boldsymbol{x}'\right)\right|\right|^2 \, d\boldsymbol{x}' \tag{2.68}$$

Similarly, Log demons uses

$$R\left(\boldsymbol{\phi}\right) = \frac{1}{\sigma^2}\int_{\Omega} \left|\left|\nabla \mathfrak{log}\, \boldsymbol{\phi}\left(\boldsymbol{x}'\right)\right|\right|^2 \, d\boldsymbol{x}' \tag{2.69}$$

In the case of diffeomorphic demons, equation (2.68) does not actually guarantee a diffeomorphic transformation. Diffeomorphism is guaranteed by the optimization method, which is explained later. This guarantee is provided by Exponential and the group properties of diffeomorphisms. As was demonstrated in section 2.3.2, the Exponential of any VVF results in a diffeomorphic DVF, so Log demons is inherently diffeomorphic.

LDDMM takes a different approach. However, it is similar to the Log domain approach in that it uses integration to ensure that the transformation is diffeomorphic. The transformation in LDDMM is parameterized by a VVF that varies in time.

$$\dot{\boldsymbol{\phi}}\left(\boldsymbol{x},t\right) = \boldsymbol{v}\left(\boldsymbol{x},t\right) \tag{2.70}$$

$$\boldsymbol{\phi}\left(\boldsymbol{x},t\right) = \int_0^t \boldsymbol{v}\left(\boldsymbol{x},t'\right) \circ \boldsymbol{\phi}\left(\boldsymbol{x},t'\right)\ dt' \tag{2.71}$$

This is in contrast to the Exponential mapping defined in section 2.3.2 where the VVF does not vary with time. The regularization term is applied such that the VVF is smooth.

$$R\left(\boldsymbol{\phi}\right) = \int_0^1 \int_\Omega \left|\left|L\boldsymbol{v}\left(\boldsymbol{x}',t'\right)\right|\right|^2\ d\boldsymbol{x}'\ dt' \tag{2.72}$$

where L is a differential operator. The differential operator is usually taken to be $L = \alpha\nabla^2 + \beta\nabla\left(\nabla\cdot\right) + \gamma$, where the Laplacian is the vector Laplacian taken over each component of the velocity vector field separately. The first term penalizes roughness in the velocity; the second term penalizes divergence (local compression or expansion); and the third term penalizes large deformations. This regularizer is inspired by fluid mechanics and describes a viscous compressible fluid. The benefit of this formulation is in its fluid interpretation since the regularization is applied infinitesimally along the flow. Like a viscous fluid flowing into the sharp corners of a container, the fluid regularizer can "relax" into regions where the response of the linear operator would otherwise be large. The regularizer in (2.68) acts on the transformation as a whole, and this relaxation does not occur.

Optimization

Once the regularizer and image distance terms have been selected many choices can be made with respect to actually minimizing the cost function. Image registration methods considered here can be summarized by the following iterative process:

1. Transform the moving image to the fixed image's coordinates
2. Compute the gradient of the image distance term
3. Compute a step direction from the gradient
4. Regularize and update the transformation

This is repeated until convergence, and is often called an alternating optimization approach [69]. Steps 1 and 2 are the same for the three methods discussed. The differences come from the step direction and regularization and update step. Technically, LDDMM requires optimization over the entire time-varying VVF. However, that approach is computationally expensive. Typically, a *greedy* approach is used. With a greedy method, the locally optimal step is taken at each iteration. With the greedy approach, LDDMM fits into the process described in this paragraph.

The remainder of this section is organized as follows. First, regularization is discussed, followed by the computation of a step direction. Finally, issues related to the transformation update are addressed.

There are two distinct places where regularization can be performed. First, the image update term can be regularized. This is called fluid-like registration because it allows fluid-like relaxation as discussed above, and, aptly, this is where LDDMM regularizes. Second, the transformation (or Log domain representation) can be regularized following the actual update. This is called diffusion-like regularization and is where demons-type algorithms traditionally regularize. More modern implementations of demons-like algorithms typically regularize in both places.

Regularization is a smoothing process and is performed by solving a partial differential equation (PDE). With LDDMM, this is explicit in the regularization term. However, the demons algorithms both regularize with convolution with a Gaussian. The Gaussian is the fundamental solution to the heat equation, which minimizes equation (2.68), so this is an acceptable operation. Greater regularization is provided with a wider convolution kernel. With LDDMM, greater regularization is provided by increasing the constants α, β, and γ.

Fundamentally, these methods are gradient descent methods. Given the gradient of the image distance functions for the images and the current estimate of the transformation, a step direction is selected to minimize the cost function, and a step is taken in that direction. Simple gradient descent is the method employed by greedy LDDMM. A fixed step size is selected, the gradient is scaled by that step size, and a step is taken in the opposite direction of the scaled gradient (subject to issues of update described below). Alternatively, the demons methods use a Gauss-Newton optimization to alter the direction of the step. This typically results in faster convergence.

The previous topics in this section are familiar from many topics in image analysis. Define a cost function, determine the gradient, and use the gradient to minimize the cost function. Typically, this means taking the current estimate and adding an update to it. However, recall from section 2.3.2 that diffeomorphisms do not form a group under addition. Instead, they are a group under functional composition. Keeping with the definition of a group, if both the initial estimate of the transformation is diffeomorphic, each update step is diffeomorphic, and the estimate and update are composed, the resulting transformation will also be a diffeomorphism. Again, each of the three methods discussed here use different strategies to ensure this.

Any sufficiently small DVF is a diffeomorphism. By integrating the compositions as in equation (2.71), LDDMM the resulting transformation is guaranteed to be diffeomorphic. In the discrete implementation, the update step is $\boldsymbol{\phi}\left(\boldsymbol{x}, t+\Delta t\right) = \left(\Delta t \boldsymbol{v}\left(\boldsymbol{x}, t\right)\right) \circ \boldsymbol{\phi}\left(\boldsymbol{x}, t\right)$; however, a sufficiently small Δt must be chosen to ensure that the update is sufficiently small as to remain diffeomorphic itself. This disallows the Gauss-Newton methods used by the demons methods. To overcome this, the demons methods use the Exponential mapping. The update step is Exponentiated before being composed with the current estimate $\boldsymbol{\phi}_{k+1} = \mathfrak{exp}\, \boldsymbol{u}_k \circ \boldsymbol{\phi}_k$. The difference between diffeomorphic demons and Log demons is that Log demons returns a transformation in the Log domain. $\boldsymbol{v}_{k+1} = \mathfrak{log}\left[\mathfrak{exp}\, \boldsymbol{u}_k \circ \mathfrak{exp}\, \boldsymbol{v}_k\right]$. This is implemented using BCH to avoid taking Logs of diffeomorphisms. This Log domain allows additional improvements to Log demons, which are explained in the next section.

Symmetric Log Demons

Symmetric Log demons is an extension to Log demons that ensures that $\boldsymbol{\phi}_{I\to J} = \boldsymbol{\phi}_{J\to I}^{-1}$. This is done by solving

$$\hat{\boldsymbol{\phi}} = \arg\min_{\phi} \frac{1}{2}\left[d\left(I\left(\boldsymbol{x}\right), J\left(\boldsymbol{\phi}\circ\boldsymbol{x}\right)\right) + d\left(J\left(\boldsymbol{x}\right), I\left(\boldsymbol{\phi}^{-1}\circ\boldsymbol{x}\right)\right)\right] + R\left(\boldsymbol{\phi}\right) \tag{2.73}$$

This is enabled by the Log-Euclidean framework. As stated in section 2.3.2, $\boldsymbol{\phi}^{-1} = \mathfrak{exp}\left(-\boldsymbol{v}\right)$, which provides the inverse at virtually no cost. The update step is computed in both directions, and linearity in the Log domain means that both updates can be combined with simple projection. That is, the update steps are used to update the transformation with

$$\boldsymbol{v}_{k+1}\left(\boldsymbol{x}\right) = \frac{1}{2} G_{\text{diff}} * \left[\text{BCH}\left(G_{\text{fluid}} * \boldsymbol{u}_{\text{for}}, \boldsymbol{v}_k\right) - \text{BCH}\left(G_{\text{fluid}} * \boldsymbol{u}_{\text{back}}, -\boldsymbol{v}_k\right)\right] \tag{2.74}$$

where $\boldsymbol{u}_{\text{for}}$ is the update step for $I \to J$ and $\boldsymbol{u}_{\text{back}}$ is the update step for $J \to I$. In [70], this symmetric method is shown to outperform diffeomorphic demons and has the benefit of increasing the meaningfulness of the inverse transformation by negation of the Logarithm, which is used in this work to transfer anatomical information within the patient.

2.4.2 Group-wise Registration

Pair-wise non-rigid registration finds a transformation that places one image into the coordinate system of another. Often studies examine a population of many subjects and desire to place these subjects in a common coordinate system where each voxel in the common coordinate systems corresponds to the same anatomy in each of the images in the population. It is possible to select a template image to provide the coordinate system and register the population to the template. However, the choice of template and changes in appearance within the population will bias the results. The goal of group-wise registration is to find an image that is *central* to the population in some way. This notion of central is captured by the Fréchet mean, which can be extended to the group-wise registration problem.

The simple linear mean for images that have undergone transformation will be blurry and is not an adequate summary of the population, and such a summary would ignore the transformations that are of interest in, for example, studying anatomic variation. The following equation states the group-wise registration problem with a Log-Euclidean metric:

$$\overline{I} = \underset{\overline{I}', \{\boldsymbol{\phi}_i\}}{\arg\min} \sum_{i=1}^{N} \left[d\left(\overline{I}'\left(\boldsymbol{x}\right), I_i\left(\boldsymbol{\phi}_i \circ \boldsymbol{x}\right) \right) + R\left(\boldsymbol{\phi}_i\right) \right] + \left\| \sum_{i=1}^{N} \mathfrak{log}\, \boldsymbol{\phi}_i \right\|^2 \tag{2.75}$$

where $\overline{I}$ is the mean of the transformed images, each I_i is an image in the population, and the final term penalizes the distance of the mean transformation from the identity.

The group-wise registration problem includes an intensity distance from a mean image and a transformation distance from a common coordinate system. The mean image is the image that requires the smallest transformation and intensity change in order to describe the population. Solving this equation is much the same as described in section 2.4.1, but instead of having a single, constant fixed image, the fixed image evolves over iterations. With an initial choice of transformations for each member in the population, the population images are transformed to the common coordinate system, and the linear mean of the transformed images is taken to be $\overline{I}_0$, where $\overline{I}_0$ is the estimate of the mean image at iteration 0. A single iteration of a registration algorithm is performed, registering each member of the population to $\overline{I}_0$. $\overline{I}_0$ is updated to $\overline{I}_1$, the linear mean of the transformed images given the updated $\boldsymbol{\phi}_i$'s. Of course, there are other methods for group-wise registration. Some of these will be discussed in chapter 5.

The results of group-wise registration are a set of transformations $\boldsymbol{\phi}_i$ from each population image I_i to a common space and an atlas image $\overline{I}$ in that common space. These represent an atlas that relates the points in the Fréchet mean to corresponding points in the population. This correspondence allows additional information to be transferred among the population in an unbiased manner. This additional information can include, for example, segmentations, functional, material and mechanical properties, or radiation dose. In this dissertation, the transformations indicate the deformation space of the population. The observed transformations in a population provide a data set from which a reduced dimensionality representation of the likely deformations that the population are likely to undergo can be learned.

2.4.3 Dimensionality Reduction on Transformations

Dimensionality reduction is a commonly performed statistical technique where, given a number of samples of a high dimensional random variable, a representation of that data is found such that the data can be represented with fewer variables. One of the most commonly used dimensionality reduction methods is principal component analysis (PCA). PCA makes the assumption that the random variable is drawn from a multi-variate Gaussian distribution. PCA returns an empirical mean and a set of orthogonal basis vectors (also, modes of variation or principal components) in order of decreasing variance. For PCA to be effective, much of the variation should be contained in the first components. It is then expected that the final few bases explain little variation in the data set and can be discarded with limited loss of fidelity to the original data set. A graphical example of PCA is shown in figure 2.14.

PCA is calculated by first subtracting the empirical mean $\boldsymbol{\mu}$ of the data set. The samples are organized into an $n \times m$ matrix $\boldsymbol{X}$ where n is the dimensionality of the data and m is the number of samples. A covariance matrix is calculated with the outer product $\boldsymbol{\Sigma} = \frac{1}{m}\boldsymbol{X}\boldsymbol{X}^T$. $\boldsymbol{\Sigma}$ is positive semi-definite. The eigenvector problem is solved to diagonalize $\boldsymbol{\Sigma}$ with $\boldsymbol{V}^{-1}\boldsymbol{\Sigma}\boldsymbol{V} = \operatorname{diag}\boldsymbol{\lambda}$. Each eigenvector is a principal component. The eigenvalues and eigenvectors are permuted to order the eigenvalues in descending order. Eigenvectors with small or null eigenvalues are discarded. Sufficient eigenvectors are retained to explain most of the variance, typically 90 or 95%. The data can then be approximated with

$$\boldsymbol{X}_j \approx \boldsymbol{\mu} + \sum_i \alpha_i \sqrt{\lambda_i}\boldsymbol{V}_i \tag{2.76}$$

In this dissertation, dimensionality reduction is performed on diffeomorphisms in

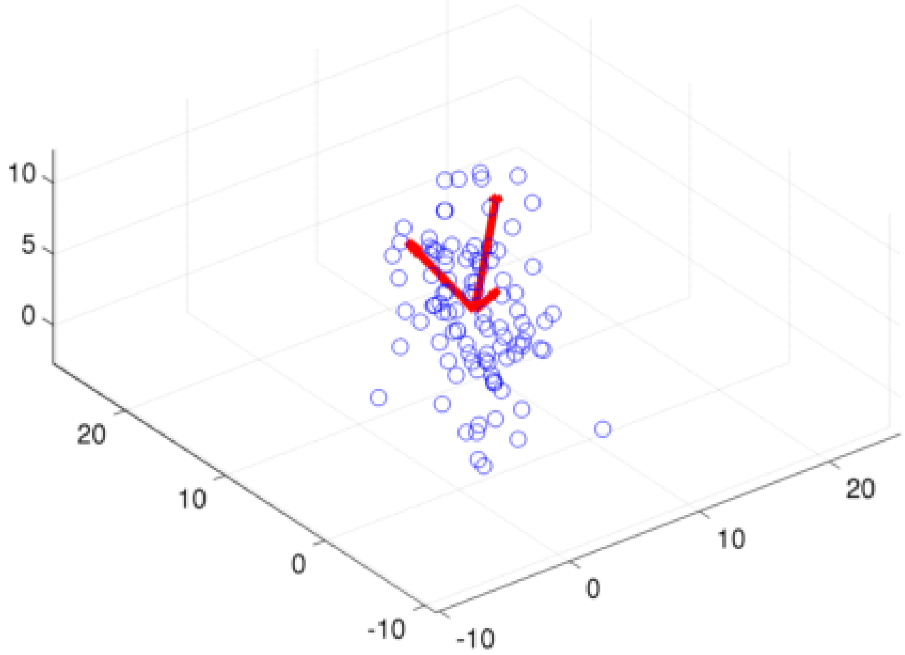


Figure 2.14: PCA provides an optimal orthogonal basis for representing a data set. The figure shows a set of data points in 3D with vectors indicating the principal directions of this basis, centered at the mean. These directions are the principal modes of variation in the data set, and their length indicates the amount of variation in that direction. By discarding the smallest modes of variation, the data set can be represented in fewer dimensions while maintaining as much fidelity to the original data as possible.

.

order to limit the solution space from that of all diffeomorphisms to a much smaller subspace consisting of more feasible transformations. PCA has been used in this goal with good results [14, 46]. However, diffeomorphisms do not form a group under addition and scalar multiplication, so, even if the principal components themselves are diffeomorphic, the space accessible by the PCA bases is not constrained to be diffeomorphic. Using the Exponential map described in section 2.3.2, an adaptation of PCA called principal geodesic analysis (PGA) can be developed that does ensure diffeomorphism [24, 9].

PGA is performed by placing a tangent space at the mean (or removing the mean from the samples, placing the tangent plane at identity, and pre-applying it to the result), projecting the samples onto the tangent plane through Log mapping, and performing PCA. Samples from the PCA subspace are then constructed normally, followed by an Exponential mapping. In this manner, the result is guaranteed to be a member of the group. For the purpose of this work, PCA and PGA are used interchangeably except where explicitly relevant.

2.4.4 3D/2D Registration

3D/2D registration can refer to many techniques, for example, slice to volume registration. In this dissertation, 3D/2D registration is limited to the registration of CT images to cone beam projections with limited angular coverage even down to single projections. The main difference between 3D/3D registration and 3D/2D registration is that the image distance function is computed on sets of 2D measured and estimated projections rather than fixed and moving 3D images. Thus, the registration incorporates the geometry of the imaging system provided by the $\boldsymbol{C}$ matrix, generalized to the projection operation $\mathcal{P}$ here, into the registration problem. In section 2.1.3, linear accelerator gantry mounted devices that can acquire these projections for IGRT were described. Simulated example images from these devices were shown in figures 2.10 and 2.11 in sections 2.2.4. These methods provide fast, low dose imaging concurrent with radiation therapy, but, due to their limited angular sampling, do not result in sufficient image quality for standard registration methods to provide adequate performance.

Using LA-CBCT it is possible to acquire a nearly complete cone beam image. Since image quality is lost gracefully with a decrease in angular sampling, 3D/2D registration problems exist along a continuum from a single X-ray projection to a complete cone beam acquisition. The amount of angular sampling, and to an extent the density of angular sampling, determines the position along the continuum, and this position motivates the solution. Projections from nearly complete angular samples can generally be reconstructed and treated as a 3D image. On the other hand, no one would consider it reasonable to reconstruct from a single X-ray projection or from two orthogonal images, although it is entirely possible. In these cases, some variant of 3D/2D registration is the only logical choice. With the angular sampling that is provided by the NST device and fast acquisitions from LA-CBCT, it is not immediately obvious whether the data should be treated more like a 3D image or more like a set of 2D projections. Indeed, the images themselves can be valuable for human interpretation. For acquisitions like this, either 3D/2D registration can be attempted or a modified 3D/3D method can be developed.

In the following paragraph, only rigid registrations between a patient's planning CT and treatment-time images are considered. The first problem with 3D/3D registration is the development of an image distance function that is adequately insensitive to the artifacts present in the limited angle reconstructions. A solution to this involves simulating the reconstruction process on a transformed version of the planning CT and was first proposed in [60]. This is known as iterative reprojection-reconstruction registration (IRRR), and takes the form:

$$\hat{\boldsymbol{R}} = \arg\min_{\boldsymbol{R}} d\left(\mathcal{R}\left(\boldsymbol{p}\right), \mathcal{R}\left(\mathcal{P}\left(I\left(\boldsymbol{Rx}\right)\right)\right)\right) \tag{2.77}$$

Here, $\boldsymbol{p}$ is a vector of measured projections, $\mathcal{R}$ is a reconstruction operator, and $\mathcal{P}$ is a projection operator, equivalent to applying the $\boldsymbol{C}$ matrix. The equivalent 3D/2D formulation is:

$$\hat{\boldsymbol{R}} = \arg\min_{\boldsymbol{R}} d\left(\boldsymbol{p}, \left(\mathcal{P}\left(I\left(\boldsymbol{Rx}\right)\right)\right)\right) \tag{2.78}$$

In [26], simulation results showed that these methods provide equivalent results under rigid transformations that can be expected during IGRT. The only difference between these methods being that IRRR takes an order of magnitude longer than 3D/2D due to the iterative application of the reconstruction operator. Since [26] focused on NST, $\mathcal{R}$ was an iterative method. It is possible that the faster FBP could be used for certain geometries, but this will still require an extra step that is unnecessary for the 3D/2D method. Since IGRT is time critical, such extra computation should not be performed unless sufficient benefit can be derived from it.

For 3D/2D registration and IRRR, there must be careful consideration in the choice of image distance terms. SSD is typically not very successful, primarily because planning CTs used for registration will be reported in Hounsfield units and the conversion factor to attenuation will be unknown. There may also be X-ray energy differences between the imagers, which leads to measured attenuation differences. Although, it would be more desirable if this could be avoided. NCC can be successful but cannot account for scatter and any non-linear intensity changes, such as those from energy differences. Mutual information is often successful [76] but is more suited for 3D/3D registrations with different energies. With rigid registrations, much of the information is obtained from bony anatomy, which is high contrast with respect to the soft tissue signal. There will always be non-rigid changes in soft tissue, and successful image distance functions will be insensitive to these. The gradient correlation method [73] performs normalized cross correlation on the gradient of the measured and estimated projections. Here gradients in soft tissue regions are small when compared with those from bony tissues. A variant of GW-LNCC is variance-weighted local normalized cross correlation [41]. Here, each window is weighted with its variance. Since the variance is small in soft tissue regions and larger when the window contains bony and soft tissue, this metric has a similar effect as gradient correlation. For non-rigid registration, where the contributions of soft tissue are more significant, GW-LNCC is an excellent choice. Other factors that are not directly addressed by the image distance terms mentioned above mainly include limitations in the planning CT. Primarily, the planning CT is considered to be the true attenuation distribution of the patient. The CT will instead be a noisy, sampled version of the true distribution. It includes artifacts, which will be reflected in the estimated projections even though they should not be, and is typically acquired with a $1 \times 1 \times 3$ mm spacing. This anisotropic spacing can increase registration error in the SI direction and is particularly challenging because the direction of poorest resolution in the CT will always be in the preferred viewing plane of the treatment-time image.

In addition to the low contrast in soft tissues, which undergo non-rigid deformation, non-rigid 3D/2D registration presents an additional challenge. For a 6 dimensional rigid registration, all transformations sufficiently close to the identity are reasonable transformations. For registrations with transformations of high dimensionality, almost all transformations are unreasonable, meaning that it is not feasible for them to be observed over the course of IGRT. Several attempts to implement an unconstrained 3D/2D non-rigid registration were made including both IRRR and 3D/2D with NST and $\sim 20^\circ$ arc

LA-CBCT, but adequate results were never obtained. However, success was reported in [61] with much larger angular coverage. Non-rigid 3D/2D registration fails for the same reason that reconstruction fails with limited angular sampling; there are many possible solutions that minimize the criteria, and the solution directly suggested by the driving forces is not the desired one. As the amount of information provided to the registration algorithm decreases, the importance of the regularizer increases.

However, rigid registration is successful. This is likely due to the fact that the transformation space only suggests feasible transformations. A possible solution to this was discussed in section 2.4.3. By performing dimensionality reduction on the transformations obtained from a group-wise registration, a low dimensional transformation space can be obtained by a linear combination of basis modes of variation. This forms a transformation which can then be applied to the Fréchet mean image and serves as a very strong registration regularizer. The few parameters of this transformation can be determined by 3D/2D registration at treatment-time, and since the transformation is constrained by the modes of variation, no further regularization is required.

The two dimensionality reduction methods described above both describe different transformation spaces. PGA with the Log-Euclidean manifold has the benefit of always producing a diffeomorphism, but it does not have a gradient that can be efficiently evaluated. In addition to the computation of the Exponential, additional computation is required for finite difference approximations to the cost function or gradient-free optimization methods must be used. The PCA model has a gradient that is simple to compute but does not necessarily result in a diffeomorphism. The impact of this choice is discussed in a later chapter.

Existing work in this area has focused on regions affected by the respiratory cycle. In these regions, the deformation a patient undergoes during treatment is largely an intra-fractional product of this cycle. Assuming that respiratory patterns are relatively constant over time a intra-patient model can be generated for every patient using respiratory-correlated CT (RCCT). In regions such as the male pelvis, deformation cannot be inferred in such a way. Male pelvic deformations are largely derived from rectal and bladder contents, changes which do not occur on a time scale that can be captured in a single imaging session.

The male pelvis has several detrimental features which make it more challenging as a target region for limited angle methods. These features are in common with the head and neck and the abdomen as target regions. Foremost amongst these is the low contrast of the structures of most interest. Under X-ray, soft tissue has very small differences in attenuation coefficient. This makes registration and segmentation difficult, and, due to artifacts from limited angle reconstructions, these differences are almost entirely obscured. When the signal is integrated in projection space (as in 3D/2D registration), structures of interest *are* invisible in individual projections. Contrast this with chest X-rays and bones. High contrast lung borders and larger nodules are visible in individual projections, as well as bones seen at all target sites. The lung also has regular and relatively simple (in terms of linear representation) deformation modes that can be well approximated knowing the location of high contrast lung structures. In [46], Liu used these to provide prior information about deformation of the lung. In the male pelvis, no such sentinel structures exist. Furthermore, the modes of deformation here are so numerous that an adequate analogous model cannot be produced with even 16 images. The major purpose of this dissertation is to develop a model which can successfully predict deformation in the male pelvis for the purposes of 3D/2D registration.

Chapter 3

Method Overview

3.1 Partitioning of Variation

A first step in the development of a deformation model (indeed, any non-rigid registration procedure) is rigid registration. To contrast with the much more challenging, high-dimensionality non-rigid registration problem, rigid registration can be performed robustly and quickly, making it a low cost step to improve model quality. Initializing non-rigid registration methods with a rigid transformation serves a dual purpose in the development of a deformation model. First, rigid transformations in medical imaging typically come in two parts: one having to do with the transformation between different device coordinate systems and another having to do with changes in patient pose. Planning-time and treatment-time images are typically acquired on different devices which will (for practical or conventional reasons) treat image data differently. For example, treatment-time image data typically exists in a treatment device specific coordinate system that is often linked to the device's isocenter (the central point around which a gantry rotates and toward which radiation is directed). An FBCT does not have an isocenter in this same sense. A translation between these can be on the order of many centimeters, and patients may even be rotated (e.g., from a head-first to a feet-first orientation). Additionally, patient setup can never be perfect, leading to pose differences between the images. This cumulative transformation can be very large, and no non-rigid registration method can be expected to recover very large transformations. Rigid registration removes these transformations from the image data, making the transformation to be recovered by the non-rigid registration smaller and, therefore, more likely to produce a desired result.

The second purpose of rigid registration is specific to the development of PCA deformation models. If rigid initialization is not performed, that rigid transformation will be reflected in the deformation fields used for model development, meaning that it will appear in a resulting model's modes of variation. This can add a significant amount of variation to the model, making it less specific and less useful. Moreover, such a PCA model does not efficiently encode the rigid registration information that it contains. Even a PCA model intended only to explain rigid transformations cannot do so as efficiently (with as few parameters) as an explicitly rigid model, and, in the rigid plus non-rigid case, modes of variation will not typically respect the independence of rigid and non-rigid modes, which can be desirable if not entirely true in practice. An optimal rigid model exists that does not require any statistical inference. It should be used.

This is not to say that statistics on rigid transformations are uninteresting. If the device dependent transformation can be separated, statistics on daily pose variations specify likely patient setup error, which can be used, for example, to modify margins in treatment plans for increased accuracy or to improve patient setup techniques. Additionally, rigid transformations form the basis of the poly-rigid transformation used in this work. This transformation, as employed in this work, can admit unrealistic articulations. However, the rigid transformations that it employs could be statistically constrained to prevent such cases.

The final three deformation models, skin, prostate-bladder-rectum organ complex (PBR), and residual, used in this work serve to partition the variation from a single deformation model into independent models based on anatomical knowledge that can be extracted by segmentation from training images. These models, in the current state of the art, can only be determined statistically. However, between rigid transformation and these statistical models, the poly-rigid transformation exists as an additional, non-statistical method that makes use of segmented training images. This poly-rigid transformation explains variation related to the articulation of the pelvis and femurs with several rigid transformations.

3.2 Poly-rigid Transformation

The poly-rigid transformation used in this work represents an articulated transformation in terms of a rigid transformation for each piece of bony anatomy and a spatially varying weight function describing the degree to which each rigid transformation affects the surrounding tissue. The weighting function is determined entirely at planning-time, and only the individual transformations for each bone are determined at treatment-time. Since bones are high contrast and rigid, it often easy to recover small transformations of them, even from limited angle images. With a properly selected weight function and properly formulated succeeding deformation models, it can be assured that rigid regions remain rigid regardless of the specific model parameters. This is in contrast to a standard PCA model where this anatomical constraint will almost never be satisfied.

Parallel to the rigid case, articulated motion that is not explained prior to the development of a statistical deformation model will be found in that model, most likely encoded less efficiently that it could be using a poly-rigid model, and a poly-rigid initialization makes the residual transformation smaller and, thus, more likely to produce a desirable result.

The poly-rigid transformation is constructed by spatially combining transformations in the Log domain, which can ensure the resulting transformation is diffeomorphic. There are two aspects to this transformation. One is the bone-to-bone rigid correspondence in that each bone has a rigid transformation to the corresponding bone in all the other images. Two is the selection of a weighting function that ensures that the rigid transformation associated with each bone rigidly maps that bone to the corresponding bone in every image and that approximates the soft tissue deformation which will occur based on those rigid transformations. The poly-rigid deformation for rigid objects indexed by b with associated weight function $w_b(\boldsymbol{x})$ and rigid transformation $\boldsymbol{R}_b$ is

$$\boldsymbol{T}_{\text{pr}}(\boldsymbol{x}) = \mathfrak{exp}\left[\sum_b w_b(\boldsymbol{x})\,\mathfrak{log}\,\boldsymbol{R}_b\boldsymbol{x}\right] \tag{3.1}$$

where $\mathfrak{exp}$ is the Exponential for $\mathbb{DIFF}^3$ and $\mathfrak{log}$ is the Logarithm for $\mathbb{SE}^3$. In this work, $w_b(x) = 1$ for $b = i$ and $w_b(x) = 0$ for $b \neq i$ in bony regions, but, strictly speaking, the transformation is locally rigid in any region where the weight function is constant.

It is obvious that this technique is not limited to forming poly-rigid transformations. Using the Exp and Log maps defined for each specific transformation group, other poly-transformations can be developed. Any set of transformations can be combined in this manner. For example, if a tumor region is expected to shrink over the course of treatment, a similarity transformation could be mixed with the rigid transformations to efficiently encode this knowledge in the transformation. Additionally, multiple deformation models could be combined to form a poly-diffeomorphic transformation.

3.3 Mean Centering of Transformations

Each of the methods here depend on the determination of an atlas image that is central to the population in some sense. The Log-Euclidean Fréchet mean is this, and it is produced by

$$\overline{\boldsymbol{x}} = \arg\min_{\hat{\boldsymbol{x}}} \sum_i ||\mathfrak{log}\ [\boldsymbol{T}_i \circ \hat{\boldsymbol{x}}]||^2 \tag{3.2}$$

where $\boldsymbol{T}_i$ is the transformation from the atlas coordinate system $\hat{\boldsymbol{x}}$ to the coordinates of the i^{th} image. If we substitute the definition of the poly-rigid transformation into equation (3.2) with w_b independent of i, then

$$\overline{\boldsymbol{x}} = \arg\min_{\hat{\boldsymbol{x}}} \sum_i \left|\left|\sum_b w_b\left(\hat{\boldsymbol{x}}\right) \mathfrak{log}\, \boldsymbol{R}_{bi}\hat{\boldsymbol{x}}\right|\right|^2 \tag{3.3}$$

where $\boldsymbol{R}_{bi}$ is the rigid transformation to the b^{th} bone in the i^{th} image. Equation (3.3) is minimized when

$$\mathbf{0} = \sum_i ||\mathfrak{log}\ [\boldsymbol{R}_{bi}\hat{\boldsymbol{x}}]||^2 \tag{3.4}$$

for each b. Therefore, the position of each bone in the atlas coordinate system is the mean pose of each piece of bony anatomy. This is independent of the choice of weighting function. The actual choice of the weighting function is discussed in chapter 4.

The poly-rigid atlas is chosen in two steps. In the first, a template image is selected and rigid correspondence is determined among the bones in the population of images. In the second, an atlas coordinate system is constructed based on the mean pose of each bone. In the template image, each of the bones of interest is segmented. These segmentations are used as masks to independently register that bone with the corresponding bone in each of the population images. This leaves a set of transformations between the template, where the transformation for each bone is the identity, and each of the remaining population images. The mean pose for each bone is then determined for each bone independently as follows.

Unlike in a Euclidean space, the space of rigid transformations $\mathbb{SE}^3$ is curved, and, using the Log-Euclidean approach, computing the mean requires an iterative method. The template defines an initial tangent point on the manifold because the transformation from the initial estimate of the atlas coordinate system to the template is the identity. Using the Logarithm at this point, each of the transformations is mapped into the tangent space. In that Log space, the Euclidean mean is taken. This mean, when mapped back into $\mathbb{SE}^3$ defines a new estimate of the Log-Euclidean mean. The inverse of this estimate is pre-applied to the population, centering them about the improved estimate of the mean. This process continues until convergence when the Log-Euclidean mean $\frac{1}{N}\sum_i \mathfrak{log}\,\boldsymbol{R}_i$ is sufficiently close to $\mathbf{0}$, making the centering transformation close to the identity. This is shown as follows:

$$\begin{aligned}\hat{\boldsymbol{R}} &= \mathfrak{exp}\left[\frac{1}{N}\sum_i \mathfrak{log}\,\boldsymbol{R}_i^k\right]\\ \boldsymbol{R}_i^{k+1} &= \boldsymbol{R}_i^k \hat{\boldsymbol{R}}^{-1}\end{aligned} \tag{3.5}$$

As noted in section 2.3.2, there are efficient, closed-form solutions to the Log and Exp mappings for $\mathbb{SE}^3$, making this very fast.

The same considerations that apply to poly-rigid transformations also apply to diffeomorphisms: the atlas coordinate system should be central to the population. The group-wise registration algorithm handles this, in some sense, by estimating a mean image, registering all the images to that mean image, and improving the mean image estimate. However, the mean determined from group-wise registration, as shown in equation (2.75), also has dependence on the both the actual images and the regularization. The mean determined by equation (2.75) is not the Log-Euclidean mean. This mean is certainly not nonsensical. It is perfectly reasonable to define a mean image that is dependent on image content, but the Log-Euclidean framework, as applied here, contains no notion of image content. The statistics computed depend solely on deformations. To remove this dependence on image content, the procedure in (3.5) is extended to diffeomorphisms in general to explicitly determine the mean and transform find the transformations from the true mean to each of the training images. An alternative is to solve (2.75) subject to the constraint that the mean image be at the Log-Euclidean mean $\sum_i \boldsymbol{\phi}_i = \mathbf{0}$. The constrained approach did not converge as well as the unconstrained, mean centering approach, and the unconstrained actually improves convergence as described below.

Since computing the Logarithm of diffeomorphisms is costly, computing Logarithms is avoided entirely. Recall from section 2.3.2 that $\mathrm{BCH}\left(\boldsymbol{a}, \boldsymbol{b}\right) = \mathfrak{log}\left[\mathfrak{exp}\, \boldsymbol{a} \cdot \mathfrak{exp}\, \boldsymbol{b}\right]$ with computation of the Log. Using this approach for diffeomorphisms was first proposed in [9]. Equation (3.5) can then be modified:

$$
\begin{aligned}
\hat{\boldsymbol{\phi}} &= \left[\frac{1}{N}\sum_i \boldsymbol{\phi}_i^k\right] \\
\boldsymbol{\phi}_i^{k+1} &= \mathrm{BCH}\left(\boldsymbol{\phi}_i^k, -\hat{\boldsymbol{\phi}}\right) \qquad (3.6)
\end{aligned}
$$

In practice, this algorithm can become unstable in less smooth regions. In order to solve this problem, Gaussian convolution with small σ was used to regularize the $\hat{\boldsymbol{\phi}}$ updates at each iteration.

However, the BCH is an infinite series of increasingly nested Lie Brackets. Because of the necessary truncation and finite difference approximation to the Jacobian, iteration of equation (3.6) introduces error into the $\boldsymbol{\phi}_i$'s, changing the correspondence determined by the registration. That is, following mean centering of the population, the image distance term is increased. In order to avoid this error, the population was registered and then re-centered in an expectation-maximization (EM) type approach. This has the added benefit of improving convergence of the registration. In all tested cases, both the image distance and the distance from the registration mean to the Log-Euclidean mean were decreased from those determined by the non-EM approach. The method also converges quickly, typically requiring only two iterations.

3.4 Summary of Transformations

The methods developed here take a set of segmented daily images from a patient and use them to develop a low dimensionality deformation model which can explain daily variations in patient setup and in deformation of the prostate, bladder, and rectum such that the parameters of that model can be determined from limited angle images at treatment-time. This is accomplished by partitioning that variation into a series of models derived from statistics and models derived from mechanics. These models are assumed to be independent and are based on anatomical information derived from segmentations. Furthermore, these transformations are constrained to be diffeomorphic by the Log-Euclidean framework, which provides machinery for efficient representation of the PCA models in the Log space, that is, PGA models.

To develop the complete model, the training images are registered together using the transformation for each stage. The set of rigidly aligned images are first registered using an articulated, poly-rigid transformation, which was described the in the previous section. The bones are aligned. A weighting function is chosen to interpolate the transformation outside of the bony regions and ensure that bony regions remain rigid. This defines the poly-rigid atlas. The images are transformed into these coordinates for further registration stages. Since the bones are already well aligned, those regions are constrained to be stationary for the remainder of the procedure. The next two stages, skin and PBR, could be performed in any order, but, for reasons explained later, I have chosen to perform the skin registration first. The skin segmentations from the daily images are transformed into the poly-rigid atlas coordinates and non-rigidly registered together, as binary images, to form skin atlas coordinates. Following this stage, the skin is aligned, and the region outside the skin (as well as the bones) is constrained to be stationary. The prostate, bladder, and rectum segmentations are transformed to the skin atlas coordinates. The same registration method is applied to the PBR stage. The final stage is the residual stage. In this stage, the bones, skin, and PBR are already properly aligned and held stationary. The images are then transformed into the PBR atlas coordinates, and the registration recovers any remaining deformations. The training images and their segmentations are then mapped into the final atlas coordinates, creating an atlas image and consensus segmentation.

PCA is the performed on the Log-domain transformations (i.e., PGA), resulting in $N - 1$ modes of variation for N images. For each non-rigid model, a subset of these that captures an adequate fraction of the total variation within the model is selected as non-rigid modes. In this work, that fraction is taken to be 95%. The the number of parameters in the cumulative transformation is then 6 rigid, 3 bones by 6 poly-rigid, and as many as are necessary for each of the three deformation models.

The order of the transformations applied is shown in figure 3.1. In each of the registration stages, transformations are determined that map the daily images into each stage's atlas coordinates. To apply the models, however, the atlas coordinates are to be transformed to daily coordinates, and image data flows up through the figure, rather than down as in model development. This requires the computation of an inverse. Because of the Log domain representation, computing the inverse of each of the transformations is cheap. $(\mathfrak{exp}\,\boldsymbol{a})^{-1} = \mathfrak{exp}\ -\boldsymbol{a}$. The inverse of each of the transformations are composed in reverse order, and the cumulative inverse transformation is applied to the atlas image.

The multiple statistical deformation model method has a weakness that is not overcome in this work. The atlas exists at some point on the Log-Euclidean manifold. At

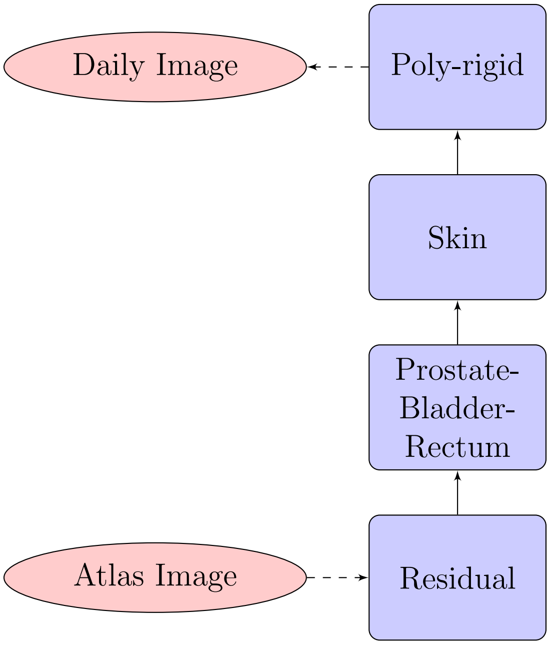


Figure 3.1: Order of transformations. The model is constructed going from daily images to atlas image, but application of this method requires transforming from atlas to daily image. Therefore, taking the inverse of the model transformation is required.

that point, there exists a tangent plane in which a PCA model has been constructed. Parameters of that model are selected, specifying a deformed image which corresponds to a new point on the manifold. The PCA model for the next stage is then placed in the tangent plane at that new point. The weakness is that specifying different parameters for the first model results in a different point on the manifold, which, because the Log-Euclidean manifold is not flat, has a different tangent plane. The same model is not strictly valid for all tangent planes.

There are two arguments as to why this approximation is acceptable. The argument discussed first is that the transformations considered here are small enough that all tangent planes can be approximated by a single tangent plane and that the manner in which the models are constructed estimates this well. The second argument discussed comes from an intuitive understanding of how the transformations are applied and how points in the atlas move as the deformations are applied.

If the space in which the transformations are considered were Euclidean (i.e., simply DVFs rather than Log-domain velocity vector fields), then the method used here is completely valid. It is that the transformations are considered in a curved space that raises these issues. Consider a method similar to the one developed in this work where transformations are points on a sphere. If the atlas is located at the north pole and the PCA model is constructed in that tangent plane, when the length of the geodesic in that plane is small, any of the new tangent planes are approximately coplanar with the tangent plane at the north pole. Due to the intra-subject nature of the transformations, a similar smallness is assumed here.

Continuing this analogy, the transformations for each of the images (after being rigidly and poly-rigidly aligned) are regarded as being at the north pole (identity). The points are then moved in the manifold such that their Log-Euclidean mean is at the identity and the geodesic from the identity to each of the points minimizes the image match and regularity terms. This process is then repeated for each of the stages. Because the samples are assumed to be at the identity when the next stage is constructed, model estimation results in a kind of averaging that reduces the error associated with this methodological weakness.

The intuitive argument for this depends on the locality of transformations (defined by construction) and the stationarity of anatomy which has already been aligned. Recall that the Log-Euclidean framework regards the Log domain representation of a diffeomorphism as a stationary velocity vector field $\boldsymbol{v}\left(\boldsymbol{x}\right)$ and that taking the Exponential is the solution to $\dot{\boldsymbol{\phi}}\left(x\right) = \boldsymbol{v}\left(\boldsymbol{x}\right)$ at $t = 1$. Points in the atlas coordinates flow through the velocity vector fields for each transformation for unit time. In regions where $\boldsymbol{v}\left(\boldsymbol{x}\right) = \mathbf{0}$, the point is stationary. The skin and the PBR are distant enough from each other that each of their respective models are zero in regions where the other model is non-zero. Since they do not interact with each other, there is no change in correspondence dependent on the parameters of the model, and it is not necessary to change the models. This non-interaction is implicit from a main assumption of this work, that each of the transformations is independent. The non-interaction also makes their ordering arbitrary. Both, however, do affect nearby tissue that has been acted on by the residual model, but, again, the falloff is rapid and most tissue is not affected by either model.

Chapter 4

Polyrigid Deformations from Bony Anatomy

4.1 Poly-rigid Transformations

Poly-rigid transformations are an attempt to spatially combine multiple rigid transformations and stem from an acknowledgment in the medical imaging community that many geometric models do not behave linearly. To aid in properly analyzing these models, the Log-Euclidean framework, which is utilized by poly-rigid transformations, makes use of the Lie Group properties of the diffeormorphism group and its sub-groups (see 2.3.2). By using the Exponential and Logarithmic mappings for the groups of interest, transformations can be mapped from their natural non-Euclidean space to a tangent Euclidean space and vice-versa, and standard Euclidean statistical methods that preserve beneficial group properties can be developed in this space.

Poly-rigid transformations involve a spatially varying combination of rotations and translations that forms a non-rigid transformation. Here, two methods of combining these transformations are discussed, the *direct fusion* approach and the Log domain approach. Ultimately, the way to manipulate rigid transformations is in the Log domain.

Figure 4.1 shows different combinations of two rotation matrices, $\boldsymbol{R}_{-45^\circ}$ and $\boldsymbol{R}_{-45^\circ}$, by two separate methods applied to an image of an ellipse with major axis in the vertical direction. The first method is the direct fusion approach, which is simply the linear combination of DVFs (equivalently, matrices). The second method is the Log domain approach used by the poly-rigid transformation. Recall that the Exponential mapping is the solution to an ODE. Therefore, the equivalent of a DVF in the Log domain is a VVF, and the Log domain approach is linear combination of VVFs (equivalently, Logs of matrices because rotations are linear transformations in the sense that they are expressed by $\boldsymbol{y} = \boldsymbol{R}\boldsymbol{x}$ rather than $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$). This example illustrates the phenomena associated with combinations of rigid transformations and is equivalent to a direct fusion or poly-rigid transformation with a constant weight function.

The weakness of the linear approach is simply demonstrated in figure 4.1. Linear combination of DVFs does not accomplish its intended purpose since the transformation does not remain rigid. Instead it demonstrates a scaling behavior. Furthermore, as shown in figure 4.2, linear combination of DVFs can result in singularities in some cases. These problems are eliminated when combination is done in the Log domain, and the resulting transformations are as expected.

Let us further interpret figure 4.2. The black circle is the set of rotations in 2D ($\mathbb{SO}^1$). $\mathbb{SO}^1$ is isomorphic to $\mathbb{S}^1$, so the complete circle (if shown) would represent all possible rotations. The plane of the page is the 2D sub-space of 2×2 matrices accessible by simple linear combination of the two 2D rotation matrices, R_{0° and R_{90°. The distance from a point the in plane to the origin is the determinant of the transformation matrix. So all transformations in this space are 2D similarity transformations (rotation and uniform scaling). The matrix logarithm maps the points on the circle to the tangent plane at the identity (the Log domain, in cyan). Since rotations in 2D have one degree of freedom, both the tangent plane and the circle have one degree of freedom. Note that points on the tangent plane do not have the same significance in the 2D space as their position in the figure may suggest. The tangent plane is a separate $1D$ linear space with distance between two points in the tangent plane corresponding to the distance

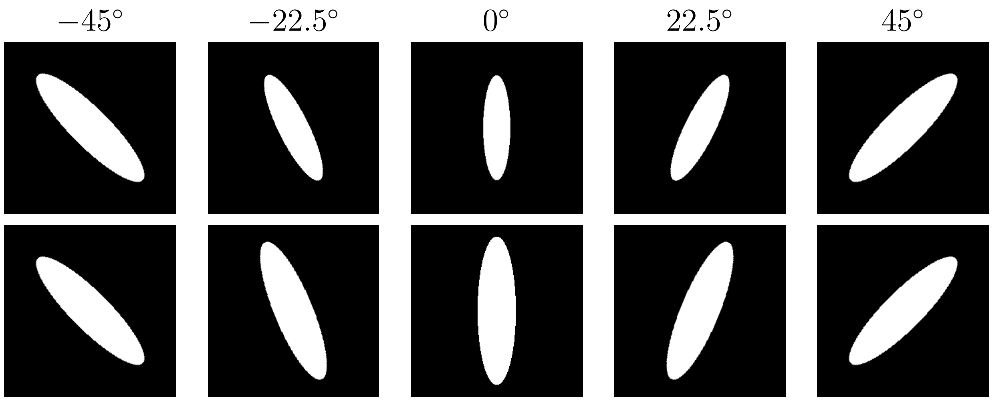


Figure 4.1: Images of an ellipse transformed by the weighted sum of a −45° and a 45° rotation matrix, $\boldsymbol{R}_{-45^\circ}$ and $\boldsymbol{R}_{-45^\circ}$, respectively, where the center of rotation is the center of the image. The combination is treated linearly (top row) and linearly in the Log domain (bottom row). Angle values are the *expected* rotation angle, with the axis of rotation pointing into the page. Note that, when treated linearly, the ellipse shrinks, meaning that linear combinations of rotation matrices are not necessarily rotations. Rotations matrices do not form a group under addition or scalar multiplication. However, when the linear combination is treated in the Log domain, the resulting matrices remain orthonormal. Figure 4.2 demonstrates the transformations geometrically.

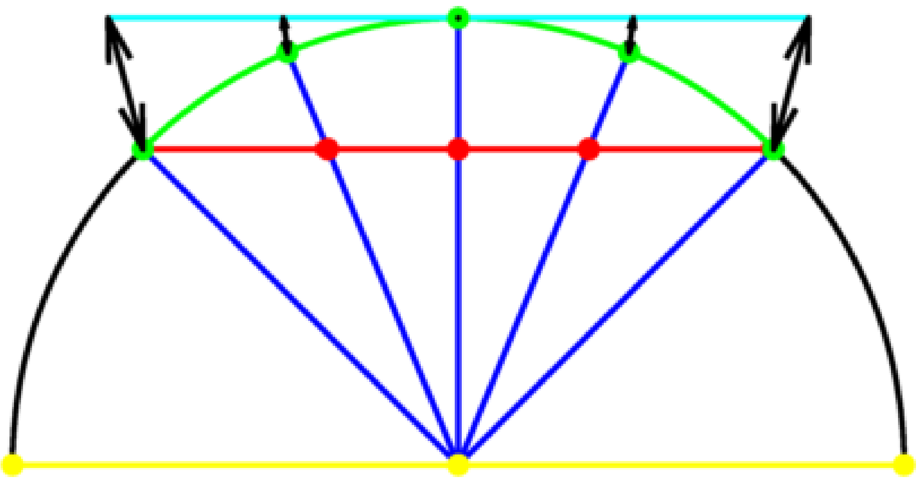


Figure 4.2: Geometric interpretation of combinations of two orthonormal matrices, $\boldsymbol{R}_{-45^\circ}$ and $\boldsymbol{R}_{45^\circ}$, both when treated linearly and linearly in the Log domain. The black line represents $\mathbb{SO}^1$. Red represents the path taken by direct fusion of matrices. Green represents the path taken when using the Log domain approach. Cyan represents the tangent plane at the identity where Log domain computations are performed. Yellow represents the path taken by direct fusion between $\boldsymbol{R}_{-90^\circ}$ and $\boldsymbol{R}_{90^\circ}$.

along the circle in the figure. The tangent plane is placed as it is to demonstrate the geometry of the tangent plane.

With the understanding from the previous paragraph, figure 4.2 can be used to interpret the transformations from figure 4.1. If rotation matrices are first mapped to the tangent plane before combination and the matrix Exponential is used to map the resulting skew-symmetric matrix back to $\mathbb{SO}^1$, then the result is always a rotation matrix. Moving a certain distance along the cyan path produces the same rotation as moving the same distance along the circle. However, if simple linear interpolation is used instead, the transformation follows the red path. Points on the red path are only rotation at the endpoints, where the red line intersects the circle. The shrinking observed in 4.1 is shown in the fact that the distance from the red path to the origin decreases. Simple trigonometry shows that the scale factor is 0.5 at $0.5\boldsymbol{R}_{-45^\circ} + 0.5\boldsymbol{R}_{45^\circ}$. The center of the circle represents the null matrix, which has zero determinant and is therefore not invertible. The yellow path passes through the origin at $0.5\boldsymbol{R}_{-90^\circ} + 0.5\boldsymbol{R}_{90^\circ}$. Therefore, simple linear combination of rotations both does not guarantee that the result will be a rotation and does not guarantee that the result be invertible.

The previous example demonstrated how poly-rigid and corresponding direct fusion transformations behave when the weight function is constant in space, but, in general, the weight function for a poly-rigid transformation will be non-constant. The previous example is extended to an actual poly-rigid transformation in figure 4.3 and is compared with a direct fusion approach. The transformations are applied to a uniform grid and shown with the log of the Jacobian determinant of the transformations. Recall that the Jacobian determinant shows the local scaling of the transformation. The weighting function in figure 4.3 ramps linearly between $\boldsymbol{R}_{-45^\circ}$ and $\boldsymbol{R}_{45^\circ}$ from left to right. The weight function sums to 1 everywhere. The weight function has a constant border to demonstrate the ability of the poly-rigid transformation to have locally rigid regions.

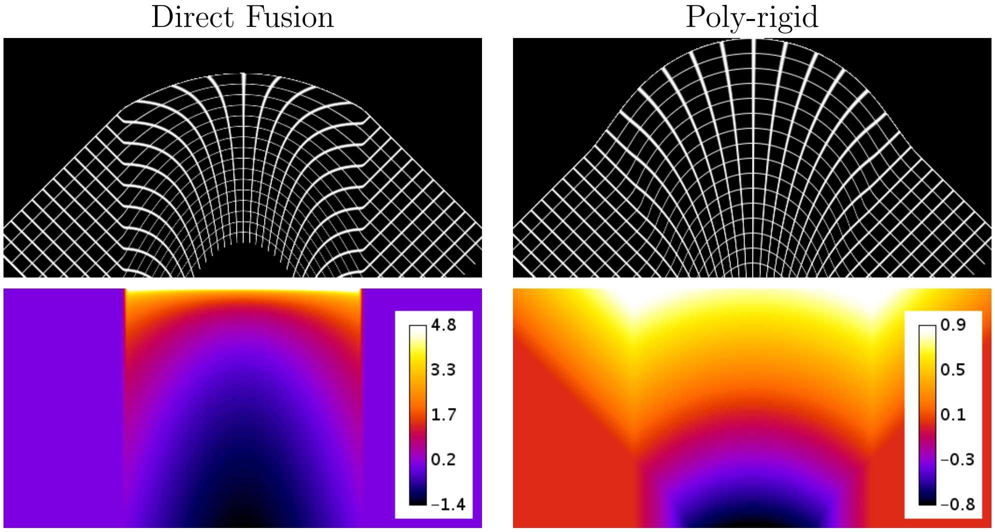


Figure 4.3: A poly-rigid and a direct DVF fusion transformation with two components $\boldsymbol{R}_{-45^\circ}$ and $\boldsymbol{R}_{45^\circ}$ with a ramp weight function that decreases the contribution of $\boldsymbol{R}_{-45^\circ}$ from left to right and increases the contribution of $\boldsymbol{R}_{45^\circ}$ to maintain a sum of 1. The direct fusion method shows the shrinking phenomenon shown in figure 4.1 while the poly-rigid method behaves as expected. Images of the log of the Jacobian determinant for the transformations are shown in the bottom row.

The direct fusion method shows similar shrinking behavior to the constant combination example. On the other hand, the poly-rigid fusion behaves as expected. At the center of the image, both $\boldsymbol{R}_{-45^\circ}$ and $\boldsymbol{R}_{45^\circ}$ contribute equally. There, the transformation is expected to be the identity. This is true for the poly-rigid transformation but not for the direct fusion transformation. At this center point, the shrinking phenomenon shown in figure 4.1 is also apparent in the direct fusion transformation. The transformations can also contain translations. Such an example is shown in 4.4. This transformation uses the same rotations and weight function as 4.3, but $\boldsymbol{R}_{45^\circ}$ is replaced with $\boldsymbol{T}_{45^\circ,-10}$, which has a translation component of -10 units in the y direction.

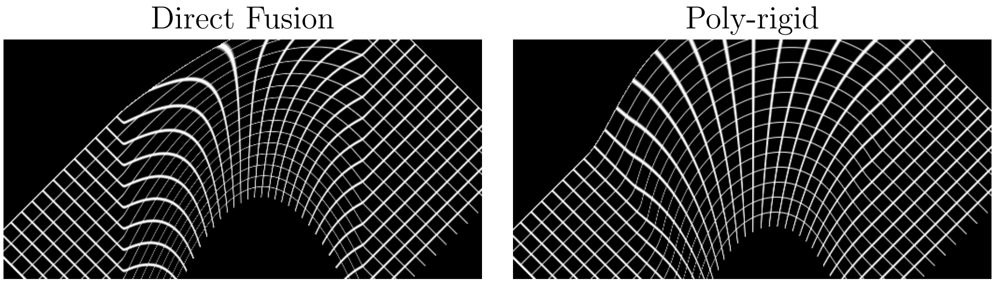


Figure 4.4: A poly-rigid and a direct DVF fusion transformation with two components $\boldsymbol{R}_{-45^\circ}$ and $\boldsymbol{T}_{45^\circ}$ with $\boldsymbol{T}_{45^\circ,-10}$ with the same weight function as 4.3.

4.2 Determining a Weight Function

This section discusses the selection of a weight function that is appropriate for maintaining rigidity in bony anatomy and approximating the effects of rigid articulated motion on tissue near to the articulating bony anatomy. The weight function is a spatially varying function that describes the contribution of each rigid transformation to a point in the domain of the poly-rigid transformation. This weight function should be selected in such a way that the bones remain rigid (that is, the weight function is constant in bony regions) and the transformations applied to tissue near to bony anatomy should be credible based on the transformation for a particular bone.

Poly-rigid and poly-affine transformations were first proposed in [3] (and extended in [4]) to provide a low dimensionality, non-linear transformation for registration of histological slices. Histological slices are prone to bending when they are sliced and mounted for imaging, so adjacent anatomical slices may not adequately align modulo a rigid or affine transformation. The author states that poly-rigid registration is particularly suited for this task because the poly-rigid transformation has the ability to remain locally rigid but still provide a certain amount of deformation. This transformation approximates the type of deformation that histological slices might undergo, and the low dimensionality parameterization helps prevent the registration from masking real anatomical differences.

However, the approach in [3, 4] and the approach in this work differ. When registering histological slices, rigid regions are not known *a priori* as is the case when using the poly-rigid transformation to align bony anatomy. In [4], both the rigid transformations and the weight functions must be optimized during the registration. In this work, the rigid transformations are determined by rigid registration in bounding boxes about each bone, and the weight function is determined solely from the shape of the bony anatomy at the mean pose (or a template pose). Because, in the registration of histological slices, weight functions are expected to be simpler, [4] uses simple Gaussians (or mixtures of Gaussians to allow more complex regions of influence) with a per-transformation relative weight to parameterize the weight function. In the histological slice application, it is also not guaranteed that there will be rigid regions, and this is compatible with Gaussian weight functions. In the bony anatomy case, rigid regions are guaranteed to be present and not well modeled by Gaussians or mixtures of Gaussians. Therefore, in this work the weight function is a general smooth function.

There are several things to consider when selecting a weight function for the application in this work. Clearly, in a bone the only contribution should be from the transformation from that bone, and near to a bone the transformation should come mostly from the transformation of that bone. The weight function should also be smooth, except at joints. Otherwise, fast changes in the contribution of each bone are not anatomically realistic. At joints, there will necessarily be a region where the weight function changes dramatically. The femoral heads and the pelvis are very close to each other. The contribution of each femoral head should drop from 1 to 0 within a few millimeters while the contribution of the pelvis must increase from 0 to 1 in the same distance. This large change, however, is not problematic because bones are mechanically connected. The respective transformations for both interacting bones must agree in that region.

There are still two issues which remain. This first is restrictions on the range of the weight function, and the second is the behavior of the weight function as the distance to bones increases. In [3, 4], $w_b \in [0, 1]$ with $\sum w_b = 1$; this set is called the normal simplex. It is not clear that this is explicitly necessary, especially in the case where both the weight function and transformations are optimized during the registration. However, in this work, image intensity information is not used, so deviation of the weight function from the normal simplex is not supported by any image evidence. Thus, in this work weights are constrained as in [3, 4].

Additionally, it is not clear what should happen far from the bones. In the hands, for example, variations in the pose of each bone can be very large, and the pose of the bones very accurately implies the deformation of the skin surface and spaces between the fingers. In contrast, in the case of radiotherapy in the male pelvis, pose variations are much smaller and do not predict deformation distant from the bones very well. For example, most deformation far from the bones will be due to fat loss and manual positioning of flesh. These types of deformation cannot be predicted from the pose of bony anatomy are not expected to be well explained by the method developed here. As such, the poly-rigid transformation used here is intended to apply best to regions closest to the bones. The other variation is expected to be explained by the succeeding deformation models. This fact decreases the importance of the specific behavior of the weight function far from the bones.

As has been discussed, the weight function used here is determined entirely from the shape of the bony anatomy, without consideration of the actual image information. However, this was not the only approach attempted. Early in method development, the weight function was determined by a registration in an attempt to address the issues of restrictions on the range of the weight function and the behavior of the weight function far from bones by looking at image information. In that formulation, the cost function for the registration was

$$\boldsymbol{T}_{\mathrm{pr},i} = \mathfrak{exp}\left[\sum_{b} w_b\left(\boldsymbol{x}\right) \mathfrak{log}\, \boldsymbol{R}_b \boldsymbol{x}\right] \tag{4.1}$$

$$\hat{I}\left(\boldsymbol{x}\right) = \frac{1}{N}\sum_{i}^{N} \boldsymbol{T}_{\mathrm{pr},i} \circ I_i\left(\boldsymbol{x}\right) \tag{4.2}$$

$$\hat{w}_b = \arg\min_{w_b} \sum_{i} \int_{\Omega} \left(\hat{I}\left(\boldsymbol{x}\right) - \boldsymbol{T}_{\mathrm{pr},i} \circ I_i\left(\boldsymbol{x}\right)\right)^2 \, d\boldsymbol{x} + \alpha \sum_{b} \int_{\Omega} \left|\left|\nabla w_b\left(\boldsymbol{x}\right)\right|\right|^2 \, d\boldsymbol{x} \tag{4.3}$$

subject to the constraint that bones are only affected by their respective transformation and with the rigid parameters initialized to their “correct” parameters, those determined from rigid registration. However, this optimization was costly, and the solution had a strong dependence on the smoothing parameter α. If α is not large, unrealistic, over-fitted weight functions are produced. For example, a pocket of tissue lateral from the left femur could be affected by the right femur. If the α is large, the solution tends to ignore the image data, and simply solves

$$\arg\min_{w_b} \sum_{b} \int_{\Omega} \left|\left|\nabla w_b\left(\boldsymbol{x}\right)\right|\right|^2 \, d\boldsymbol{x} \tag{4.4}$$

The solution to the particular optimization in (4.4) neither provides particularly good results nor adheres to our assumption that the region near to a bone should be influenced mostly by that bone. That is, the component of the weight function for a particular bone initially falls off too quickly and continues to have a global effect. Because this formulation was not successful in solving the weight range and distant behavior issues or providing an effective weight function, it was abandoned in favor of the one that follows, which only uses image information in the sense that the rigid transformations and segmentations were derived from it.

The weight functions used in this work are found by solving

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}\in W}{\arg\min} \sum_b \int_\Omega \left|\left|\nabla w_b\left(\boldsymbol{x}\right)\right|\right|^2 + \alpha D_b\left(\boldsymbol{x}\right) w_b\left(\boldsymbol{x}\right)\ d\boldsymbol{x} \tag{4.5}$$

such that all elements in $\boldsymbol{w}$ are on the normal simplex and that bones are only influenced by their respective transformation. Here, $D_b\left(\boldsymbol{x}\right)$ is some function of the Euclidean distance to each of the bones, to be described later. Gradients were determined by the standard approach, and the problem was solved by conjugate gradient descent for increased convergence speed over standard gradient descent. This formulation was selected empirically based on the ability of the resulting transformation to reduce the sum of squared differences between a template patient image and many daily images of that same patient in a region near to the bones.

Constrained optimization problems of this type are typically difficult to solve because they require an equality constraint and two inequality constraints, as well as treatment of the bony regions. Since the bony regions are fixed, handling them is trivial. Those regions are initialized appropriately and gradient is always set to zero. To handle the other constraints, the gradient was considered in the log domain with the mapping

$$w_b = \frac{\exp \omega_b}{\sum_i \exp \omega_i} \tag{4.6}$$

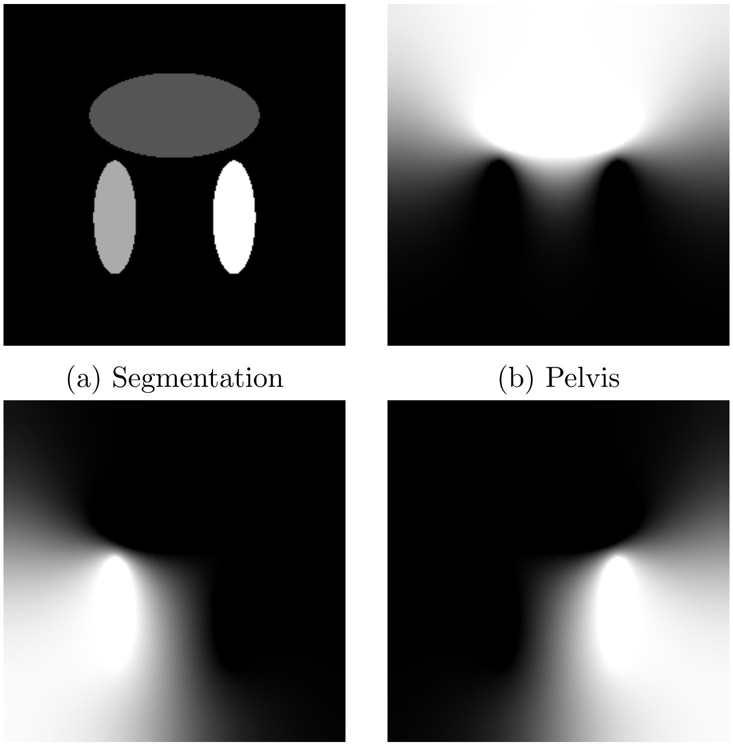
The exponential ensures that w_b is always positive, and the denominator ensures that $\sum_b w_b = 1$, which guarantees that $w_b < 1$. This approach does remove the possibility that elements can be influenced by *exactly* one transformation, but the influence of other bones can be made so small as to be practically zero.

The distance term $D_b\left(\boldsymbol{x}\right)$ is required to ensure that a bone only affect tissue near to it and that regions very close to a bone are influenced almost entirely by that bone. As described above, if (4.5) using on the bony and range constraints of $\boldsymbol{w}$ (that is, $\alpha = 0$), individual transformations can have a very global effect. Instead, α is used to ensure that bones have more local effects. For reasons described below, the function D_b used in this work is

$$D_b = \begin{cases} -\frac{1}{d_b} & \text{if bone } b \text{ is nearest bone} \\ \frac{d_b}{d_n} & \text{otherwise} \end{cases} \tag{4.7}$$

where d_b is the Euclidean distance to bone b and d_n is the Euclidean distance to the nearest bone, computed with GNU General Public License (GPL) code from [22]. The contribution of the transformation from the nearest bone is driven higher by the inverse of the distance to that bone. Contributions from the other transformations are driven down by the ratio of the distance to the respective bone to the distance to the nearest bone. This is done so that locality penalties become smaller as the distance to the bones increases, increasing the relative strength of the smoothing parameter in regions where the nearest bone changes. However, because the ratio of distances is used, the penalty becomes smaller if the bones are nearly the same distance from each other.

To illustrate the procedure, an example problem was constructed that approximates the bony anatomy of the pelvis, and poly-rigid transformations were generated from it for a sample of large, but feasible bone poses. The segmentation of the example image into simulated pelvis and left and right femur with weights determined by the method discussed in this section is shown in figure 4.5. In figure 4.6, the initial image and the image deformed using the generated weights and simulated femur pose changes of 10, 20, and 30 degrees about the tips of the ellipse which represents the bones is shown. Because the rotation is not about the origin (the center of the image), there is translation involved in this example.



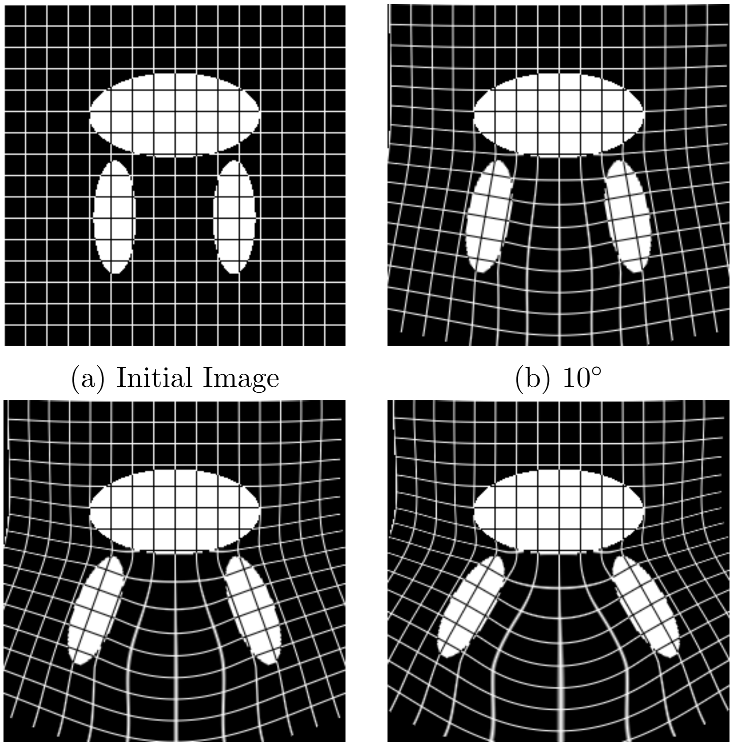
(a) Segmentation

(b) Pelvis

(c) Right Femur

(d) Left Femur

Figure 4.5: Segmentation of the example problem and the weights generated by solving equation (4.5).



(a) Initial Image

(b) 10°

(c) 20°

(d) 30°

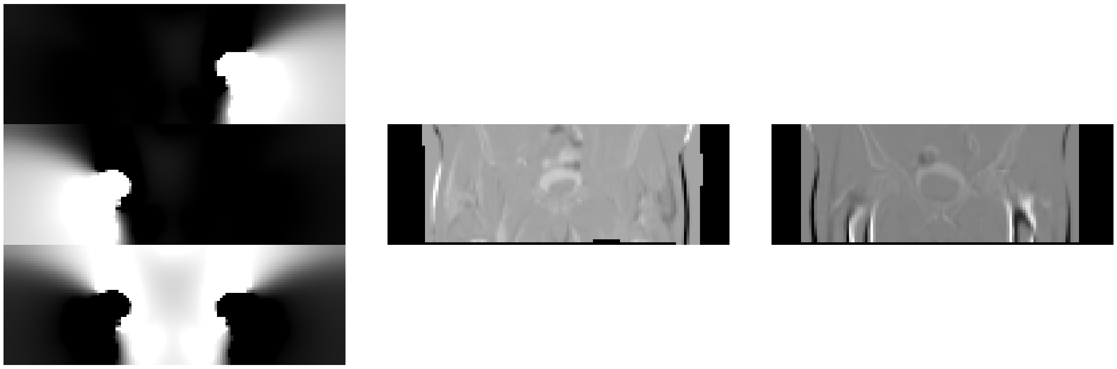
Figure 4.6: 10 degrees, 20 degrees, 30 degrees has associated translation since rotation is not about the origin

4.3 Results

The purpose of the poly-rigid transformation in the method developed in this work is to correctly and rigidly align bony anatomy so that articulated rigid motion will not appear in the subsequent deformation models that are to be developed and to provide a better initialization for those subsequent non-rigid registration stages under the assumption that better initialization to a non-rigid registration algorithm leads to better results.

Visual inspection is used to demonstrate that bony anatomy is properly aligned when using the poly-rigid transformation described in this work. Figure 4.7 shows a typical weight function along with image differences between the template image to which the moving image is registered using poly-rigid alignment and rigid alignment. The pelvis and both femurs were segmented in the template image using a threshold and mathematical morphology method. The resulting single bone mask was then manually separated into individual bones and edited for correctness. The resulting segmentation was used to generate weights according to (4.5) with parameter $\alpha = 0.075$. In order to determine the transformations for each of the bones, the images were first rigidly aligned based on SSD with a threshold applied to the images remove contributions from soft tissue. Then, each of the bones in the template was individually rigidly registered to the moving image by optimizing over the SSD calculated only in the segmented bone. The three rigid transformations and the weights were then used to generate a poly-rigid transformation according to (3.1). Both the poly-rigid and rigid transformations were applied to generate the deformed image. Visual inspection shows that the bones are properly aligned.

To quantify the effects of poly-rigid alignment of bony anatomy, the process used to generate the sample image in 4.7 was repeated for three patients with a single template



(a) Weight Function (b) Poly-rigid (c) Rigid Alone

Figure 4.7: Example poly-rigid weights and transformation applied to a patient image. b shows the difference between the template image and the poly-rigidly aligned moving image, and c shows the difference between the template image and the image aligned using rigid registration alone. The bones are much better aligned in b than in c. Because the global rigid registration tends to better align the pelvis than the femurs, the pelvis is relatively well aligned in c, but improvement can still be seen by using the poly-rigid transformation.

image and 16 daily images for each patient. The SSD for the rigidly aligned and poly-rigidly aligned images was calculated in regions within 5 cm of any bone. The relative change in image distance for all the images for each patient is shown in figure 4.8, where the relative change was calculated

$$\Delta\text{distance}_i = \frac{d_{\text{PR},i} - d_{\text{R},i}}{d_{\text{R},i}} \tag{4.8}$$

where $d_{\text{PR},i}$ is the SSD between the template image and the poly-rigidly aligned image and $d_{\text{R},i}$ is the SSD between the template image and rigidly aligned image. There was a significant decrease in image distance for all patients. However, there was no statistically significant difference ($p < 0.05$) in image distance when SSD was calculated over the entire image. This is due to large variations in skin surface which this poly-rigid transformation is not expected to handle.

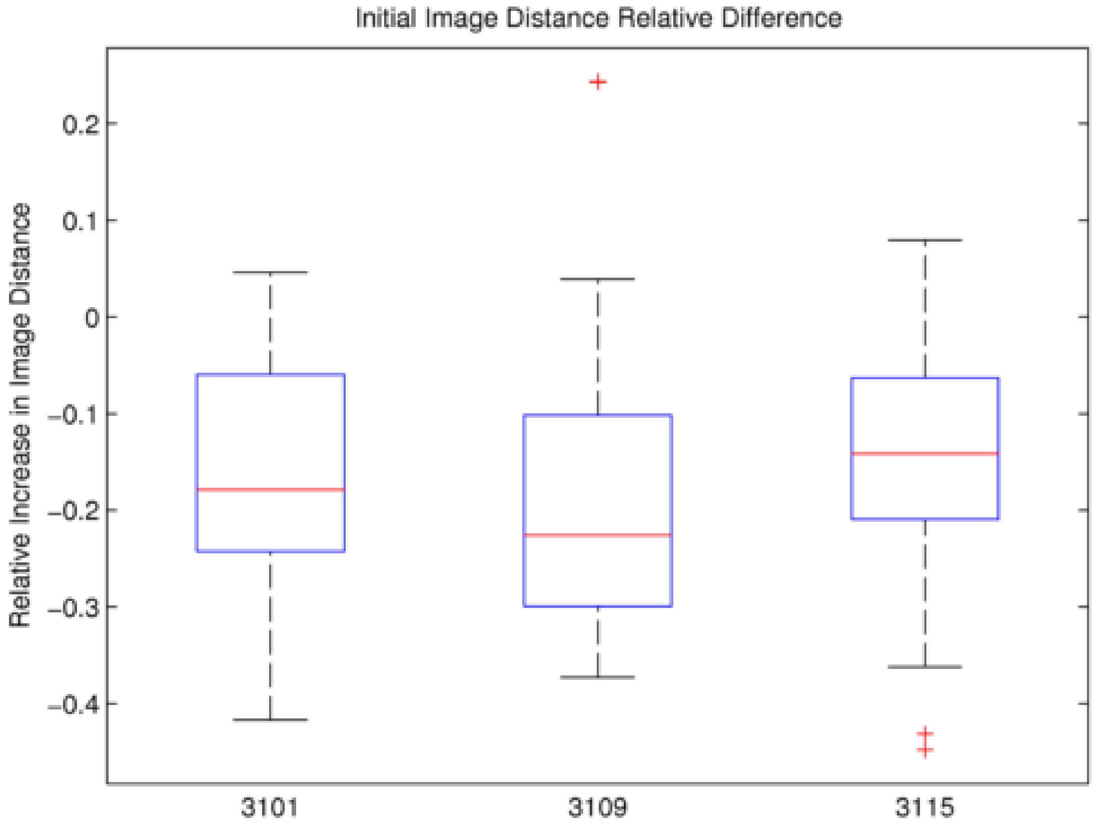


Figure 4.8: Distribution of the relative changes in SSD between rigidly aligned and poly-rigidly aligned images. Poly-rigid alignment significantly decreased image distance for all three patients.

A poly-rigid alignment also improves the quality of succeeding non-rigid registrations by providing a better initialization than rigid alignment alone. The images used to generate figure 4.8 were then further non-rigidly aligned to template image using symmetric Log demons [70] with SSD as the image distance. The distribution of relative differences in image distance after non-rigid registration is shown in figure 4.9, and the differences calculated according to equation (4.8). Using a poly-rigid initialization to a non-rigid registration provided a statistically significant decrease (pair-wise T-test, $p < 0.05$) in the image distance after the non-rigid registration for all three patients. The average improvement being 14%. Figure 4.10 visually demonstrates the differences between initializing using a poly-rigid transformation and a rigid transformation alone.

4.4 Summary and Conclusions

Chapter 4 provided increased detail about the behavior of the poly-rigid transformation and the necessity for treatment of rotations in the Log domain in order to ensure both that poly-rigid transformations behave as intended and that the resulting transformations remain invertible. The process for determining a weight function that, together with known rigid transformations for each bone, well approximates the deformation of tissue due to articulation of the pelvis and femurs was described. It was demonstrated that the proposed poly-rigid transformation accurately and rigidly aligns bone anatomy. The proposed poly-rigid transformation was used to register daily images of several patients to their template planning image. It was determined that performing a poly-rigid transformation along with a rigid transformation significantly decreases the image distance term when compared with using a rigid transformation alone. Proper alignment of the bones was demonstration by visual inspection. The poly-rigidly aligned and rigidly aligned images were then non-rigidly registered to their template images. After the non-rigid registration, the poly-rigidly aligned images retained a significantly

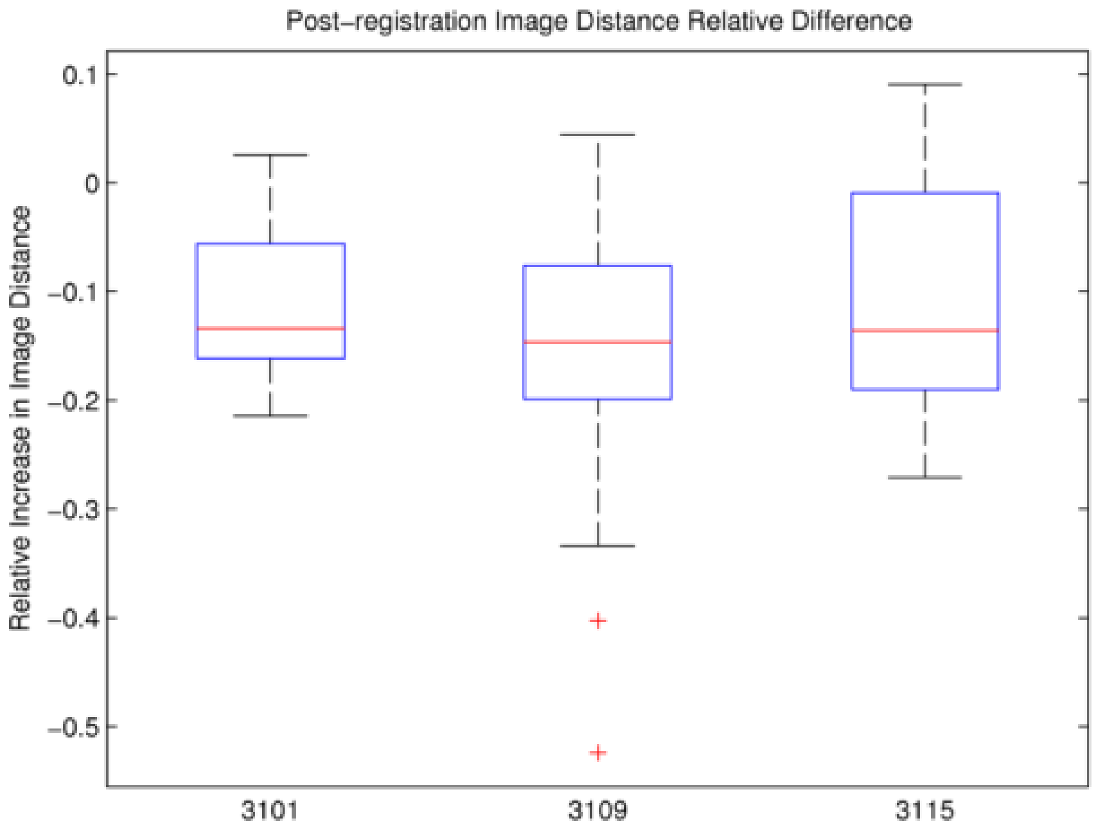
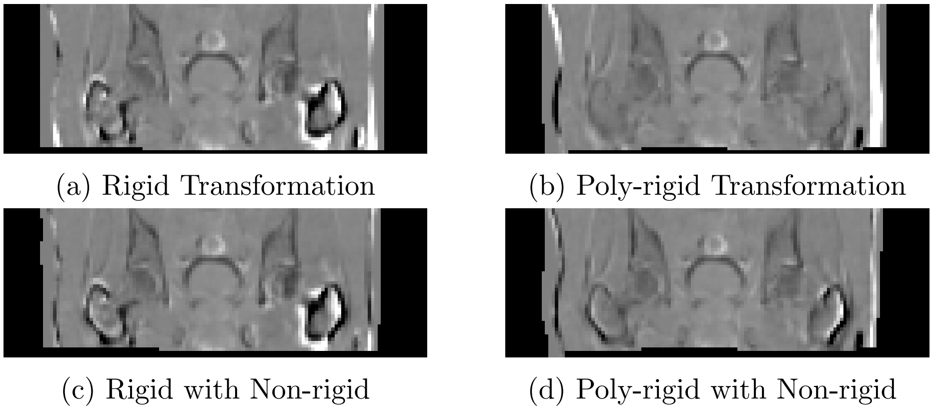


Figure 4.9: Distribution of the relative changes in SSD following a non-rigid registration which was initialized with rigidly aligned and poly-rigidly aligned images. Initializing the registration algorithm with a poly-rigidly aligned image significantly decreases the image match when compared with only rigid initialization.



(a) Rigid Transformation

(b) Poly-rigid Transformation

(c) Rigid with Non-rigid

(d) Poly-rigid with Non-rigid

Figure 4.10: Difference images between the fixed image and the images aligned by rigid transformation (a), poly-rigid transformation (b), rigid initialization to non-rigid registration (c), and poly-rigid initialization to non-rigid registration (a). Black regions are caused by padding in regions where image data is not defined. Bones do not remain well aligned in d when compared with b because they were not constrained to remain aligned. The constraint that bones remain aligned during non-rigid stages is used in this work and is further discussed in section 5.2.2

better image than the corresponding rigidly aligned images, indicating that performing a poly-rigid alignment prior to a non-rigid registration increases the quality of that registration.

The importance of the poly-rigid transformation in this work is three-fold.

First, the poly-rigid transformation allows (and, in this application, is forced to have) locally rigid regions. This enables the rigidity of bones to be preserved. If instead of accounting for articulated motion in this way, articulated motion were captured by the subsequent deformation models, the bony regions would almost never remain rigid, even if the non-rigid transformations from which the deformation model would be developed were rigid in bony regions (which is typically not the case unless the registration is constrained). Because bones have high contrast with their surrounding tissue, they have a tendency to overwhelm the signal from lower contrast soft-tissue boundaries. Bony anatomy that projects to the same region of the detector as soft tissue will entirely prevent the visualization of that soft tissue. Ensuring proper alignment of bones in order that their contribution to detector signal is appropriately accounted for is critical to the success of this method.

Second, the poly-rigid transformation reduces the number of modes of variation in subsequent non-rigid deformation models. That is, articulated motion that is explained by the poly-rigid transformation does not appear in these subsequent deformation models, meaning that fewer modes of variation are necessary to determine a sufficiently accurate deformation model. Deformation models in the male pelvis already need to explain a large amount of variation. There is sufficient variation in the male pelvis that a single PCA-based deformation model does not produce an effective shape space for registration from even as many as 16 daily images. In this work, the solution is to separate the single deformation model into several independent models, thereby reducing the amount of variation to be explained by each model. The poly-rigid transformation serves to further reduce the amount of variation that needs to be explained by those models by removing the effects of articulation.

Third, because the poly-rigid transformation has removed deformation due to articulation, the remaining deformation to be recovered by a succeeding non-rigid registration is smaller, meaning the registration is better initialized. Registrations with better initializations typically result in more accurate transformations (as demonstrated by the reduction of image distance in figure 4.9). These more accurate transformations should produce a better shape space, resulting in a more accurate deformation model.

Chapter 5

Multi-Deformation Models for Tissue Deformation

5.1 Introduction

By this point in the document, the reader will have a good understanding of the method discussed in this work. A patient-specific deformation model is constructed which consists of the following:

1. An atlas image of the patient with consensus segmentations of bones, skin, and PBR

2. A poly-rigid transformation that accounts for articulated motion of the femurs and pelvis, as well as the motion of soft tissue and muscle near to that bony anatomy

3. A skin model that accounts for deformation of the skin and regions near to the skin

4. A prostate, bladder, and rectum model that accounts for deformation of that organ complex

5. A residual model that accounts for all other changes in the patient's anatomy.

This model is constructed from daily, segmented images of a patient. The parameters of this model are determined at treatment-time based on limited angle 3D/2D registration.

These parameters provide a transformation from the atlas image and any of the patient's planning or daily images to treatment-time. This transformation provides both an approximate CT-like image that is suitable for dose accumulation and a segmentation of the patient's prostate, bladder, rectum, skin, and bony anatomy (in terms of the deformed atlas). Chapter 4 provided details about the construction of the articulated poly-rigid transformation. This chapter provides further details about the construction of the non-rigid tissue deformation models and their application at treatment-time.

5.2 Deformation Model Learning

As introduced in section 2.4.2, group-wise registration is the registration of a population to a common coordinate system, resulting in an atlas image and transformations from the common coordinate system to each image in the population. The atlas image provides a common coordinate system in which, for example, average shape and appearance can be observed. The variation in the population is apparent in the transformations. This variation can be summarized, for example, by PCA to represent likely modes of variation that the atlas image can be expected to deform. In this work, three such models are developed: skin, PBR, and residual. Each model is largely the same; it is a group-wise symmetric Log demons registration. Their similarities are discussed first, and they are contrasted later. This is interspersed with relevant information regarding the implementation of the models.

5.2.1 Group-wise Symmetric Log Demons Registration

Traditionally, group-wise registration studies suffered from a bias issue. That is, a single template had to be chosen from the population to which all images would be registered. The atlas image would be biased towards that arbitrarily chosen image. Additionally, this image would not necessarily be central to the population (as in a mean), which is a desirable property for statistical analysis. Alternative approaches were attempted where the sum of the DVFs were constrained to be $\mathbf{0}$ during the registration [6], but this approach is only successful for *small deformations*. Small deformation situations are those where deformations are small enough that the composition of transformations is sufficiently well approximated by addition of their DVFs. This makes Euclidean averaging of DVFs meaningful. Small deformations are not clearly distinguished from large deformations in the literature, and it is not clear that such a distinction can be made non-arbitrarily. This large deformation is the "large deformation" in LDDMM because LDDMM is designed to handle these large deformation situations. The Log-Euclidean framework also handles large deformations and can be constrained such that the VVFs that it returns sum to $\mathbf{0}$ (that it, a large deformation implementation of [6]). However, as discussed in 3.3, this constrained approach is not as successful as the approach used in this work, which is described later in this section.

The first attempt at unbiased, diffeomorphic, group-wise image registration that makes use of the manifold structure of diffeomorphisms was published in [38]. There, the LDDMM framework is used to register the population to a continually evolving mean. The initial mean image is the linear mean of the population images affinely registered to a common space. Population images are registered towards this initially fuzzy atlas, and the atlas is updated at each iteration based on the newly transformed images for that iteration. In [35, 9], the Log-Euclidean framework was used to reduce template selection bias. There, an initial template is selected to which all the population images are registered. A mean deformation is then computed using (3.6). The inverse of the mean deformation is pre-applied to each of the transformations, resulting in a centrally located, mean coordinate system. Each of the population images are then transformed into the mean coordinate system, and a linear mean is taken to serve as a new atlas image. The population images are then re-registered to this less biased derived template image. The process continues in an EM fashion until convergence.

As was alluded to in sections 2.4.2 and 3.3, the method used in this work is a hybrid of the methods described in the previous two paragraphs. Here, at every iteration, the images are registered to a continually evolving estimate of the mean image. This is continued until convergence, similar to [38]. Then, the Log-Euclidean framework is used to determine an explicit mean of the transformations using equation (3.6). The inverse of the mean transformation is pre-applied to each of the population transformations, placing the atlas coordinate system at the Log-Euclidean mean of the population transformations. This process is iterated until convergence. This registration method is used in determining the transformations for each of the three tissue deformation models. For the skin and PBR stages of model development, only two iterations were necessary, but three iterations proved beneficial for the residual stage.

This hybrid strategy has improved convergence over the methods described in both [38] and [35, 9]. This claim is made based on two metrics. The first is the sum of the image distances from the atlas to each of the transformed population images, where a smaller image distance means that the deformed population images better match the atlas image. This hybrid method resulted in smaller image distances than both alternative strategies. The second is the transformation distance, the magnitude of the transformation that maps the atlas coordinate system to the mean of the transformations from the atlas coordinate system to each of the original population images. This hybrid method decreases the transformation distance when compared with the method in [38]. However, because the final step of both this hybrid method and the method proposed in [35, 9] effectively sets the transformation distance to zero, the transformation distance for both this hybrid method and the [35, 9] method are equivalent.

This registration method also makes use of the symmetric extension to the Log demons method (previously discussed in section 2.4.1), which makes use of the Log-Euclidean framework to ensure that the results of the registration are symmetric with respect to the order of the inputs. With the group-wise extensions discussed above and previously, the extension to symmetric group-wise Log demons is straightforward. For each population image and at each iteration, a forward step (registering the population image to that atlas estimate) and a backward step (registering the atlas image to the population images) are performed. The results are then combined in the Log domain to determine a symmetric update.

5.2.2 Anatomical Constraints

As has been discussed in 3.4, after the application of a poly-rigid transformation the registration method described in the previous section is applied sequentially to skin segmentations, PBR segmentations, and finally the residual images, producing, at each stage, consensus segmentations and increasingly aligned anatomy. The consensus segmentations from the previous stages can then be used to constrain the deformation in future stages. That is, those segmentations imply that some regions of the image should not deform. This is necessary because even though the anatomy previously aligned is known to be well aligned, the image data available to the current registration stage may not suggest that fact. For example, when the PBR registration stage is performed, the image data are only piece-wise constant images of the PBR expert segmentations, but it is known that the bony anatomy and regions outside the skin should be stationary. This constraint also has the added benefit of ensuring that there is no variation to be explained by the deformation model in these already aligned regions. A description of the constraint procedure is described in the following paragraph.

Recall that the forward and backward update steps first involve the calculation of the gradient of the image distance term and then involve the regularization of that gradient by Gaussian convolution (fluid-like regularization, see section 2.4.1). Both the forward and backward updates are then combined, and the result is used to update the current estimate of the transformation, which is the regularized again (diffusion-like regularization). Also, note that regions marked as already aligned are not transformed in either the forward or backwards steps. Finally, recall that any region of the Log domain VVF that is zero becomes the identity when Exponentiated. Because of the averaging step used in the combination of the forward and backward steps and the BCH used in updating the current estimate of the transformation, if the velocity in paired voxels is zero, it will remain zero. So, to constrain already aligned anatomy to be stationary, the gradient in those stationary regions need only be set to zero. The regularization stages will alter these values, but this is handled by resetting the updates and the transformation to zero there following both the fluid-like and diffusion-like regularization stages.

5.2.3 Registration of Segmentations

In this work, segmentations are used four times to aid in the development of a deformation model in the male pelvis. First, they are used to aid in the removal of gas (to be discussed later). Second, they are used to determine the rigid transformations for each of the bones used in the poly-rigid transformation. These are first used to mask out regions of the image so that rigid transformations for each of the bones can be determined in isolation from nearby anatomy. They are later used in determining the weighting function for the poly-rigid transformation. Third and fourth, segmentations of the skin and PBR are used as image data for the development of skin- and PBR-specific deformation models. Additionally, bone, skin, and PBR segmentations are used to ensure that already registered anatomy remains stationary at later registration stages. This section discusses the issues concerning registering these segmentations to develop a deformation model.

Because segmentations are equivalent to binary functions, the problem of registering segmentations is the problem of registering shapes, and, as in image registration, the problem of registering shapes is the problem of finding correspondences in the population. However, shapes admit a more varied set of representations than do images, and these representations either aid in the finding of correspondences or provide them directly. These representations generally fall into four classes. The first class is landmark based representations. These are a set of corresponding points that are placed at recognizable locations in the shape. For example, the tips of the fingers and the interdigital space in a hand shape. These landmarks provide the only points of correspondence for the shapes. There are few if any recognizable landmarks in the skin and pelvic organs, so this representation is not suitable for this work. The second class is boundary representations. Boundary representations represent the shape by some representation of its boundary. These can be, for example, some type of ordered point set connected by line segments in 2D or a triangle mesh in 3D, or they can be based on some orthogonal function such as Fourier or spherical harmonics in 2D and 3D respectively. In the latter, correspondence can be established by properly parameterizing the population [28]. In the former, correspondence can be determined by sampling the surface such that some correspondence defining objective function is optimized. For example, a function that maximizes the ability of the point model to accurately represent the population while minimizing its statistical complexity, as in [12]. Similarly, the surface representation (mesh or line segment set) can be constructed such that the correspondence points are its vertices. The third class is skeletal representations. Skeletal representations parameterize shapes via a skeleton that is somehow central to the shape and a vector from this skeleton to the surface [64]. Correspondence throughout the entire object, rather than simply its boundary, is determined from position along and vector from this skeleton. The final class represents correspondence as diffeomorphisms from an atlas shape to each of the population shapes. Here, the correspondence is provided by the diffeomorphism. This diffeomorphism method is the approach used in this work for reasons to be explained later.

Given a shape representation and a method for obtaining correspondence from that representation, to use that information within this framework methods for dimensionality reduction and for determining a diffeomorphism from the mean shape to the members of the population must be developed. For the diffeomorphism method, dimensionality reductions is provided byPCA with the transformations as data, as discussed previously and further explained in 5.2.6 below. This method can also be used for any shape representation, but alternative methods may provide a more compact representation. One possible method is to perform PCA on corresponding points giving modes of variation in the point distribution. Discretely sampled skeletal representations (S-reps) provide another option in that PGA can be performed on the discrete elements that make up the model, namely, a position in space, direction to the boundary, and distance to the boundary. In this S-rep case, corresponding points can then be generated from both the mean shape and the deformed shape. Ultimately, this produces two point clouds, which can then be used to generate the necessary diffeomorphism. As a first attempt at generating the diffeomorphism, the distance between corresponding points can be substituted for the image data in any of the registration methods described in this work to produce a suitable diffeomorphism. More advanced methods include rotational flows [43] and methods based on Laplace's equation [63].

The diffeomorphism shape matching method was used mainly for its simplicity and similarity to the existing registration method. This allows the shape matching method to use the same notion of mean as is used in the image registration code (that is, the Log-Euclidean mean of diffeomorphisms). This is seen as a benefit because introducing

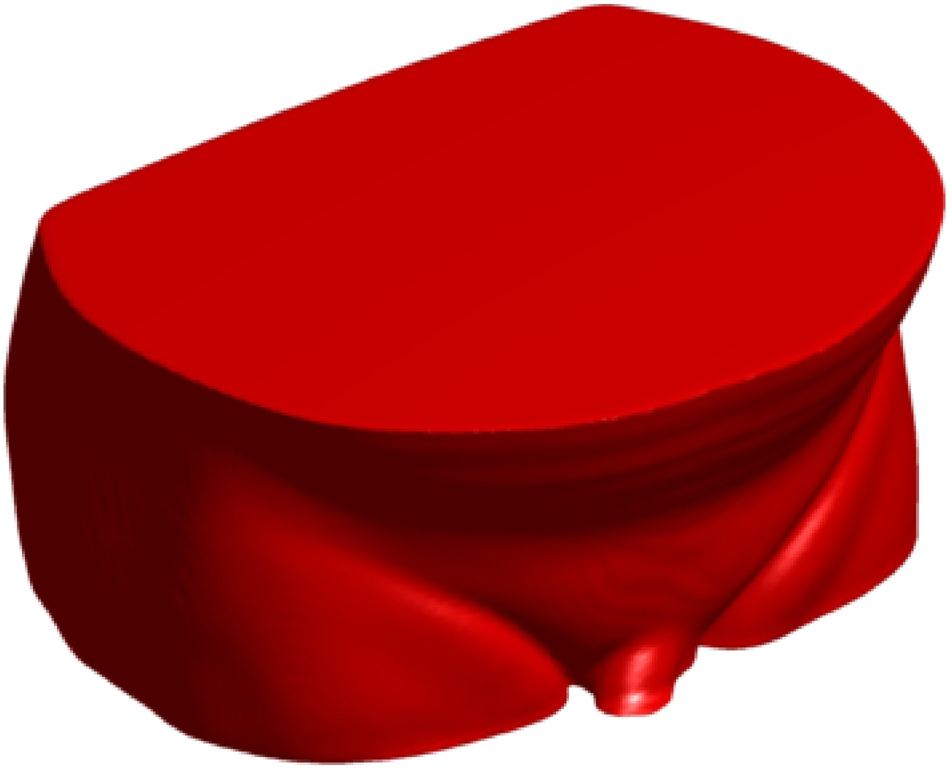


Figure 5.1: Isosurface rendering of a skin atlas.

additional notions of mean increases complexity without immediately obvious benefit.

Following rigid and poly-rigid transformations, the next registration aligns the skin. The skin-air boundary has high contrast and is easily segmented using thresholding and mathematical morphology operations. The skin segmentation is performed automatically during image preprocessing. The skin segmentation from each population image is converted to a binary image. These images are then registered together. Because the image is binary and to save time, SSD was used as the image distance metric in the registration. An example skin atlas segmentation is shown in figure 5.1. Additionally, this skin segmentation is used during image preprocessing to remove noise outside the patient and the table on which the patient is positioned.

After the skin is registered, the prostate, bladder, and rectum are aligned as the single PBR organ complex. These organs are not trivial to segment automatically,

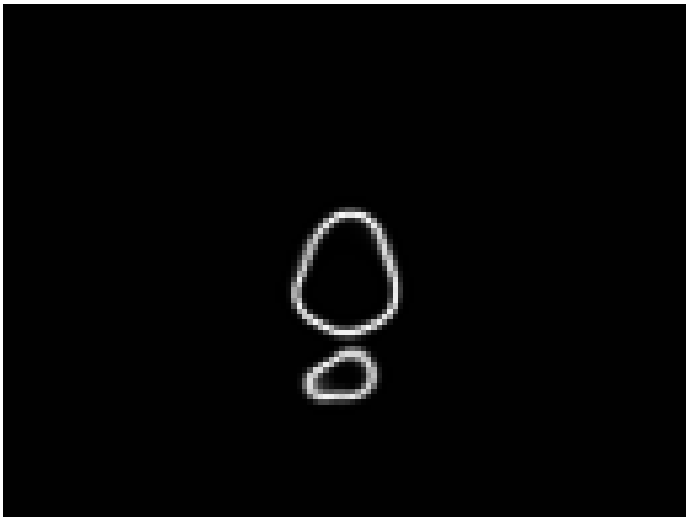


Figure 5.2: Image data for the second stage of registration during the PBR stage showing the bladder and rectum.

so manual segmentations were provided by experts. Because the organs are in close proximity to each other, each organ was assigned an arbitrary value. The choice of the value, however, can have an effect on the transformation returned by the registration. To alleviate the effects of this arbitrary choice, during the second stage of registration, when the organs are already well aligned, the edges of the organs were found, and a smoothed version of this edge image was used as image data. An example image is shown in figure 5.2. After the completion of this stage, a PBR atlas is formed, as shown in figure 5.3.

5.2.4 Residual Registration and Correction of Gas

After the skin and PBR registration stages are completed, a final stage is performed to develop a model that contains all residual variation in the patient. This is the only non-rigid registration that is performed using actual CTs as image data, all previous stages having used segmentations as image data. Because actual CTs were used, GW-LNCC was used with a Gaussian window with $\sigma = 8$mm. GW-LNCC provided visually better

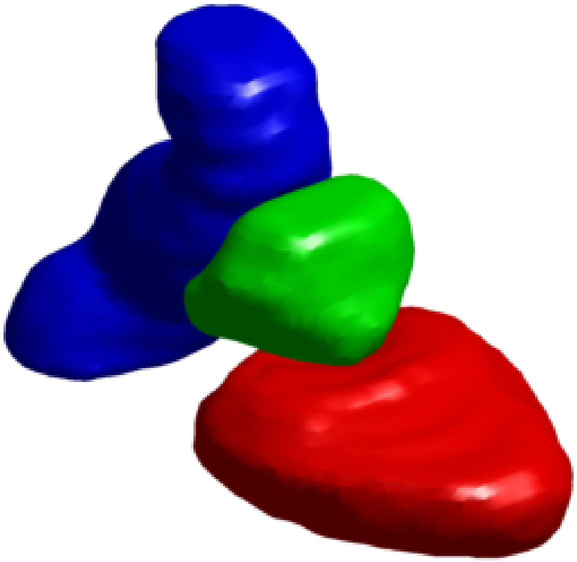


Figure 5.3: Isosurface rendering of a PBR atlas, with prostate in green, bladder in red, and rectum in blue.

atlas images (in terms of their perceived “blurriness”, where sharpness is a qualitative measurement of the quality of the group-wise registration) than did SSD alone. There are, however, questions about the *capture range* range of GW-LNCC, a larger capture range indicating a better ability of a measure of image distance to direct the registration towards the global optimum. Therefore, in the first iteration of the registration and re-centering process, SSD was used as the measure of image distance. In later iterations, GW-LNCC is used. However, neither GW-LNCC nor SSD can adequately handle transient gas bubbles in the images. This is discussed next.

A major source of error in diffeomorphic registration in the abdomen and pelvic regions is the contents of the gastrointestinal tract, particularly gas. Similarly, gas may also be found in the bladder. Bowel contents are highly variable in appearance and location, and because they are transient, there is no correct transformation that matches them. Bowel contents do, however, cause real changes in shape that ought to be accurately reflected in the shape space. However, because there is poor image match between bowel gas and other tissue and bowel contents, erroneous deformations will be produced in two ways, dependent on the presence or absence of nearby gas. First, if there is another bubble nearby, the bubble in the moving image may be drawn towards the bubble in the fixed image in an attempt to match its shape. This may overwhelm the regularizer, causing inaccurate deformation in nearby tissue. In the event that there is not another bubble near by, the gas bubble will shrink. This shrinking and associated expansion of nearby tissue will appear in the transformations and, thus, create an inaccuracy in the shape space.

Ultimately, these gas bubbles must be somehow masked from the registration. Fortunately, the same reason they cause problems to registration makes them easy to segment in preparation for correction. A typical and successful procedure is thresholding followed by mathematical morphology to eliminate small bubbles, which will have negligible effects on the transformation. In [19], gas bubbles were deflated using a method derived from LDDMM. Points at the border of the gas bubble were collapsed along their inward facing normals, and the deformations were regularized as in equation (2.72). The resulting deflated images were registered. The deflation transformation can then be composed with the transformation from the registration to approximate correspondence. This successfully masks the bubble but has the secondary effect of inducing unrealistic shape change in the deformations. As a result, it does not produce true correspondence in those regions. In this work, at least approximate correspondence is desired, so a different method is used, which is described next.

In this work, gas bubbles are filled in with tissue-like intensities. The naive approach of filling in the gas bubbles with a constant value is unsatisfactory because it introduces discontinuities in the image. Instead, this work intends to fill in the gas bubble with intensities that provide as little information as possible to the registration algorithm. These are the smoothest intensities that match the boundaries of the gas bubbles. In each of the gas bubbles,

$$\hat{I} = \arg\min_{I} \int_{\Omega} ||\nabla I||^2 \ d\boldsymbol{x} \tag{5.1}$$

is solved subject to boundary conditions that $\hat{I} = I$ at the edges of the gas bubbles. Here Ω is the domain of the gas bubble. The method is similar to Poisson image interpolation found in commercial photo editing software, where it often produces excellent results [57]. An example image can be found in figure 5.4. A similar problem occurs at treatment-time. However, no gas correction is attempted at this stage. This is because the only data available at treatment-time are limited angle 2D projections. This is discussed further in 5.3. From these, it is difficult to identify and mask gas bubbles. As such, bubbles are simply ignored. This could be a source of error in the method, particularly when there are large gas bubbles adjacent to the prostate in the rectum. In practice, this did not cause sufficient issues. The argument being that the transformations suggested by the image data would not be accessible from the shape space.

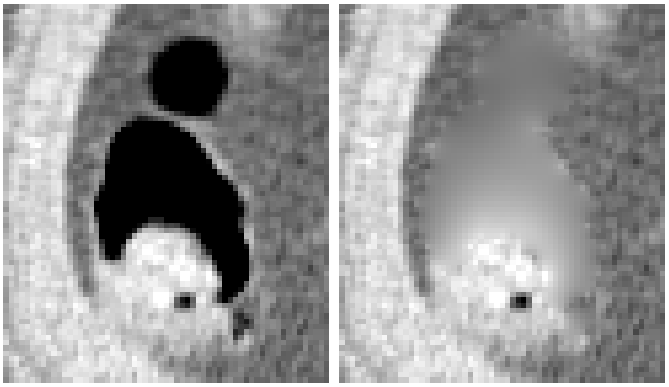


Figure 5.4: CT slice of a typical gas bubble before and after filling using the method in this work. The small bubble in the lower portion of the region was disconnected from any larger gas pockets and smaller than the volume of consideration.

After all stages of registration, a final atlas image is generated, an example of which is shown in figure 5.5.

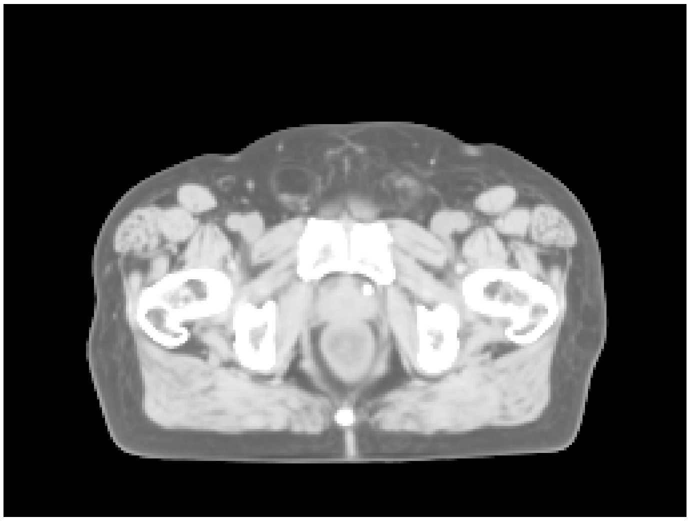


Figure 5.5: Typical, windowed atlas image for a patient, showing rectum and prostate.

5.2.5 Effects of Image Truncation During Model Learning

Because CT involves the use of potentially dangerous ionizing radiation and radiation dose is proportional to the FOV, clinical CTs are typically truncated to the smallest region that contains the anatomy to be examined. This is particularly true for the daily CTs used in some adaptive radiotherapy (ART) protocols but causes problems both in the development of deformation models, to be explained in this section, and in the application of those models at treatment-time, to be explained in a later section. Typically, daily CTs for an ART protocol are typically taken to be as small as possible while still ensuring that OARs are fully imaged. This is all that is necessary to perform segmentation and dose accumulation as in a typical ART workflow. A standard gantry-based linear accelerator only delivers radiation isocentrically. That is, beams are directed from the circular path of the beam portal towards the machine isocenter. As a result, significant dose will only be deposited in a small slab of tissue, which would roughly correspond to this small image. Because the images used to develop and test the models used in this work were obtained from such a clinical protocol, they do not have the larger FOV that would be desirable for this work.

For reasons that will be explained when this topic is revisited in section 5.3.2, a deformation model that is longer in the axial direction is more desirable. However, the axial length of a deformation model derived from a set of CTs is the minimum axial length of the intersection of all of the rigidly registered CTs because outside of this region there is not available image information from which to determine correspondence for every image. This is a common issue in registration (at least, in those registration problems where the image data cannot be reasonably preprocessed into images that are "floating in air", such as the brain). Fortunately, this is only a problem in the superior and inferior regions of the image. The left, right, anterior, and posterior extents of the patients are fully imaged, except in the case of larger patients who exceed the reconstruction volume. For these edges, the patient may be considered to be "floating in air", and the images can be padded with air values. For single image-to-image registration problems, this SI truncation is typically not a problem. The fixed image is chosen to be the domain of the problem, and, if the moving image is too small, it is padded with some signaling value (for example, NaN), which indicates that image information at that point is not available. There, the gradient with respect to the image data is not calculated, and the deformation is implied solely by nearby regions defined in both images and the action of the regularizer. This solution is not nearly as adequate in group-wise registration problems. There the domain of the problem is chosen to be larger than the union of all of the images, and all of the images are padded with a signaling value. The problem is that as image extents are reached, the spatial effects of the regularizer prematurely bias the mean space towards the images with larger extents. Depending on the application of the group-wise registration, this bias may not be too much of a concern. However, if the transformations are to be used to develop a deformation model, the problem is that in regions outside of the intersection of all the images will not have a valid set of transformations on which to perform dimensionality reduction. That is, because there is no image data to guide the registration, any regions that are not the identity are only affected by the regularizer, which does not provide a realistic transformation in this situation.

Ultimately, the only option is to solve the registration problem on a domain smaller than the size of the intersection. However, there is additional information from most or all of the images outside of this region. This again becomes important near the image extents. While it is reasonable to assume identity boundary conditions at the other image extents, this is not the case at the superior and inferior extents. In these areas, it is probable that some transformations will map out of the domain. In typical implementations of non-rigid registration algorithms, all images are rigidly registered and resampled into the solution domain, effectively throwing out data beyond the edges. This work alleviates this problem by storing images with their original FOV, just resampled, allowing transformations that map outside of the solution domain to map to real image data in most cases. Although, there still can remain situations where transformations that can map to unknown values (in this situation, the signaling value approach is used). This process attempts to minimize their effects and produce transformations that are suitable for construction of deformation models and that are as accurate as possible near these edges.

5.2.6 Dimensionality Reduction

In order to develop a deformation model from the set of transformations, dimensionality reduction is performed on the transformations to find a set of modes of deformation that the patient is likely to undergo. In this work, PCA is used (described in section

2.4.3) in two variants, which are compared in the results section later in this work. The actual dimensionality reduction method PCA is the same for both the variants. The difference in these two methods is the nature of the data on which PCA is performed. PCA makes the assumption that the modes of variation can be combined linearly. That is,

$$\boldsymbol{T} = \boldsymbol{\mu} + \sum_i \alpha_i \boldsymbol{M}_i \tag{5.2}$$

where $\boldsymbol{\mu}$ is the mean, $\boldsymbol{M}_i$'s are the modes of variation, and α_i's are the weighting factor combining each of the modes. This is not a valid assumption for diffeomorphisms, since diffeomorphisms do not form a group under either addition or scalar multiplication. Consequently, the resulting transformation is not necessarily invertible. As has been mentioned previously, this does not make the process nonsensical, particularly in the case of small $\alpha_i \boldsymbol{M}_i$'s. However, if the data are considered in the Log domain, the group properties can be preserved using the Exponential. This dimensionality reduction on data considered in the Log domain is known as PGA. This makes the construction of a transformation for the purposes of this work

$$\boldsymbol{T} = \mathfrak{exp} \sum_i \alpha_i \boldsymbol{M}'_i \tag{5.3}$$

where the $\boldsymbol{M}'_i$ are the modes of variation derived from the Log domain data. Since the transformations in the Log domain have been mean centered, the mean is negligible and is not added to the combined modes of variation.

While performing dimensionality reduction in each of the models is necessary for this method, there is an alternative place where PCA/PGA could have been performed. This is on the poses of the bones for the poly-rigid transformation. There are three bones in the model each with 6 parameters describing their rigid pose, for a total of

18 parameters. These parameters are certainly linked in some sense, and, only a small fraction of the pose space is accessible. For example, if the pelvis is rotated about the SI axis, the femurs must be transformed to accommodate this. Additionally, the left femur cannot translate very much unless the pelvis is translated accordingly.

Dimensionality reduction on these data face the same data domain, PCA/PGA choice. The standard PCA approach is obviously not the correct one for dimensionality reduction of rigid transformations, that is, dimensionality reduction on matrices reflecting rigid transformations. As was discussed in section 4.1, this does not do what is intended. PGA on the Log domain parameters is an obvious choice because it can maintain the group properties. A possible alternative approach can use Euler's angle representations of the rotation and translation. Euler's angles represent rotations as a sequence of rotations about a specific axis. However, one should not use Euler's angles, even though they guarantee rigidity, because they are non-linear, and PCA is not intended for non-linear data.

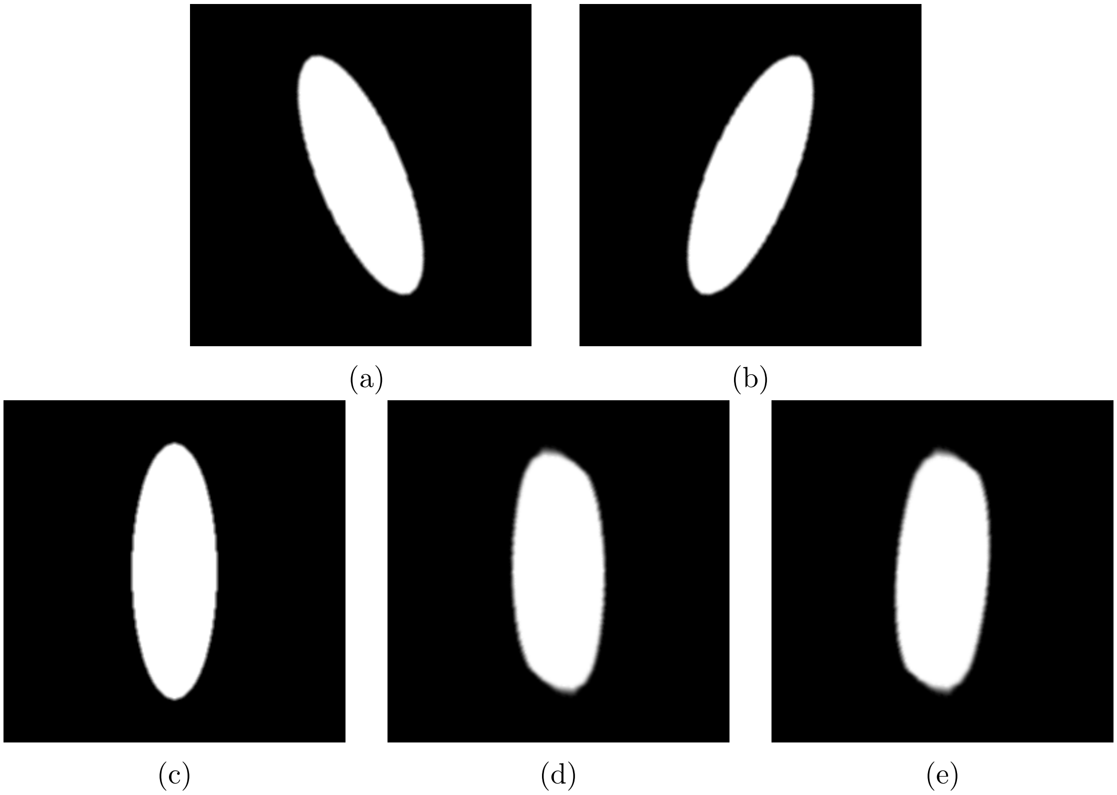
Ultimately, no dimensionality reduction was performed on the rigid poses. First, dimensionality reduction is not necessary. Bones are high contrast and are salient even in 2D projections. In this work, there was no trouble accurately recovering these poses. Second, when dimensionality reduction is used, the space of possible recovered poses is decreased. If the actual pose that the patient is in is not near to the dimensionally reduced space, the actual pose cannot be recovered without error. This error can obscure the lower contrast transformations of interest, which affect the prostate, bladder, and rectum. Because this source of error can be avoided while still accurately recovering the poses of the bones, the optimization was performed directly on all 18 rigid parameters.

5.2.7 Geodesics of Transformations

In this work, transformations are parameterized in several ways. The first is purely rigid transformations used to initially align patients for model learning and align models to daily images for model application. The second is poly-rigid deformation described in chapter 4. The third is non-rigid deformation. Both the second and the third parameterizations have two sub-parameterizations, parameterization in the linear domain and in the Log domain. In section 4.1, it was demonstrated that linear representation in the poly-rigid domain is not effective. However, it is not *a priori* obvious whether the Log domain representation of non-rigid transformations is superior to the linear representation, aside from its theoretical superiority. This issue will be discussed further in section 5.4. This section discusses some aspects about about the choices of transformation parameterization and its effects on the geodesics of transformations.

The non-rigid transformations that are investigated in this work are parameterized as in equations (5.2) and (5.3) as sums of modes of variation. It should be noted that these modes of variation are in themselves valid transformations. These modes are controlled with a single scalar value α_i making each mode a one-parameter subgroup (see 2.3.2). The question is what transformations are members of the one-parameter subgroup and are these the transformations that are desired. An example is shown in figure 5.6. In figures 5.6a and 5.6b, two ellipses are shown. There are infinitely many possible transformations between the two. If no information is known about the nature of these objects (as in the actual physical objects that the images represent), it is ultimately unclear how a valid object intermediate between the two would appear. An obvious choice is to assume rigidity. Figure 5.6c shows the intermediate ellipse midway between the two ellipses. Rigidity is known only for bony anatomy, and this rigidity has been employed to improve the method in this work. Figures 5.6d and 5.6e show two alternative intermediate images, also midway between the two ellipses. These intermediate images were derived based on determining the intermediate transformation (obtained by symmetric Log demons) in both the Log domain (5.6d) and the linear domain (5.6e). This registration did not obtain what I would consider the "most correct" transformation, that is, the one where pixels move directly in the x direction based on their distance from the center of rotation. This would obtain the same correspondence between the two images as if the transformation were rigid. This, however, furthers the notion that there cannot be a single "most correct" intermediate transformation because an alternative intermediate transformation can be imagined where the major axis of 5.6a shrinks and the minor axis expands to match 5.6b. This type of intermediate transformation is seen in the greedy implementation of LDDMM. However, the intermediate transformations found in the greedy implementation of LDDMM do not truly follow a geodesic, and I do not know if this type of intermediate transformation would appear when using the full implementation of LDDMM. As such, this example is not shown.

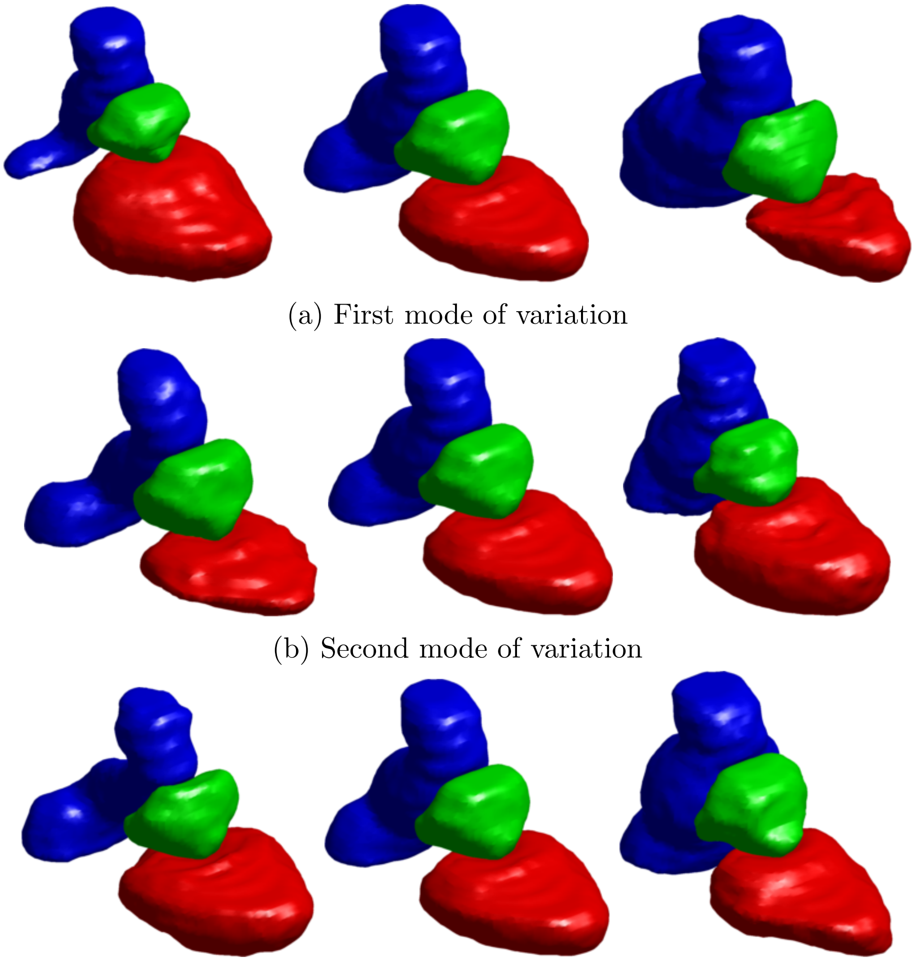
Based on the above discussion, the accessible transformations based on equations (5.2) and (5.3) produce a valid space based on the available training data but not necessarily the best or "most correct" one for the population as a whole. The modes of variation, however, are realistic. The first three modes of variation for the PBR model are shown in figures 5.7 and 5.8 for Log domain and linear domain treatment, respectively. Modes of variation for the poly-rigid transformation have been shown in section 4.3. Modes of variation for the skin and residual models are credible but do not have obvious physical interpretation (such as the organ volume changes shown in the PBR modes) and, therefore, are not shown.



(a) (b)

(c) (d) (e)

Figure 5.6: The ellipses in a and b represent a fixed and a moving image. There exists a one-parameter subgroup of transformations between the identity and a transformation mapping a to b. Figures c, d, and e show the ellipse transformed according to the transformation intermediate between the transformation mapping a to b using rigid, Log domain, and linear geodesics respectively.

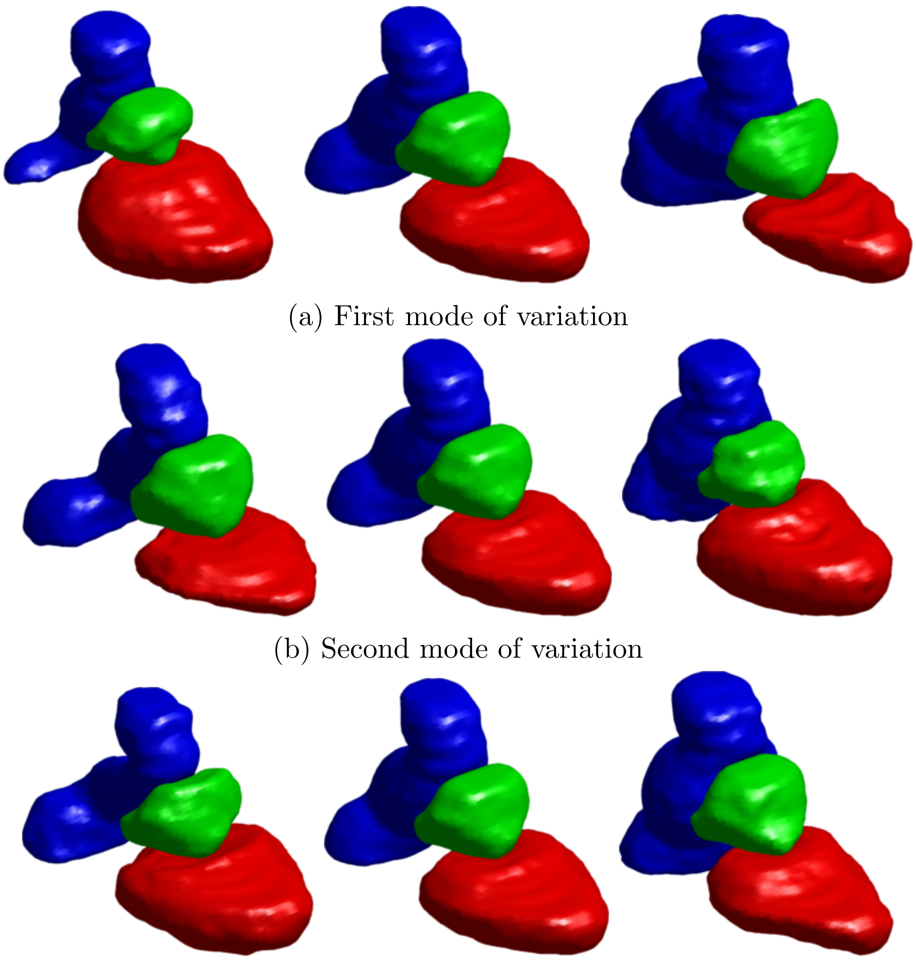


(a) First mode of variation

(b) Second mode of variation

(c) Third mode of variation

Figure 5.7: First three modes of variation using the Log domain representation in the PBR model with -3σ, mean model, $+3\sigma$ on the left, center, and right respectively. The first mode shows an expansion of the superior portion of the rectum associated with a shrinking of the bladder. The second mode shows an expansion of the medial section of the rectum below the prostate associated with an expansion of the bladder. The third mode shows mostly rectal shape change in the absence of large bladder volume changes.



(a) First mode of variation

(b) Second mode of variation

(c) Third mode of variation

Figure 5.8: First three modes of variation using the linear domain representation in the PBR model with -3σ, mean model, $+3\sigma$ on the left, center, and right respectively. Note this visual similarity to figure 5.7

5.3 Treatment-time Application

Previous sections discussed the development of a poly-rigid transformation model and the three non-rigid deformation models from 3D CT images. At treatment-time, the only data available are the 3D pre-learned model and a set of limited angle 2D X-ray projections. This section is concerned with setting the parameters of the pre-learned model from these limited angle projections. The process of determining these parameters is a technically simple 3D/2D registration. First, the atlas image is rigidly registered to the set of 2D projections. This rigidly transforms the model into the treatment-time coordinate system. An optimization[1] is then performed on the remaining poly-rigid and non-rigid parameters, matching the projection of the atlas image deformed according the parameters and the measured projections. There are, however, several complicating issues that need to be addressed.

5.3.1 Inability to Calculate the Derivative

The first issue is that there is not an efficient way to compute the derivative of the Exponential transformation with respect to its parameters. While the derivative can be computed, it is more computationally costly than using a finite difference approximation. Therefore, gradient-free optimization methods must be used. In this work, two separate optimization algorithms are used.

In the first stage, the Nelder-Mead optimization method [53] (also known as the downhill simplex or amoeba method) is used. Nelder-Mead is an ordered search method that explores the optimization space strategically evaluating the cost function. A set of candidate solutions are chosen, an initial estimate of the solution $\boldsymbol{x}_0$ and N additional candidate points $\boldsymbol{x}_i$, where N is the dimensionality of the optimization problem. These

[1] In this work, the parameters of the model are found by performing an optimization. For 3D/2D registration approaches that do not use an optimization see [16, 14, 15]

candidate points are typically chosen to be $\boldsymbol{x}_i = [x_{0,0}, x_{0,1}, \ldots x_{0,i} + \lambda_i, \ldots x_{0,N-1}]$, hence the name simplex. The value of the objective function is evaluated at each of these candidate points. The points are then adjusted in a prescribed fashion by flipping, shrinking towards the center, and expanding in order to explore the objective function in search of a local optimum. The benefit of the Nelder-Mead is that by using a large (but appropriately sized for the problem) $\boldsymbol{\lambda}$ the algorithm can be a less local optimization than gradient-based methods. This can make it more robust to non-convex objective functions. However, Nelder-Mead is very slow to converge in poorly-scaled functions, and, in fact, it need not converge at all [49].

To overcome this issue, after a certain number of iterations or a loose convergence condition is satisfied (for Nelder-Mead, a convenient termination criteria is a threshold of the hyper-volume of the simplex), a second optimization is performed where the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method [25] is used. The BFGS method is a quasi-Newton method where the successive gradients from each iteration of the algorithm are used to build up an approximation of the Hessian matrix of the objective function. However, BFGS does require knowledge of the gradient of the objective function. This is provided by the finite difference approximation. The optimization is then performed twice: first with a forward difference for speed and second with a central difference for accuracy.

5.3.2 Model Length Concerns During Application

It has been previously explained in section 5.2.5 that the deformation models learned in this work are necessarily undesirably truncated in the superior inferior direction. Because of this truncation, the detector area that the projected model covers is very small. The issue is further exacerbated by the fact that only rays that enter and exit through the anterior, posterior, left, or right model extents. This is because rays that enter or exit through the inferior or superior image extents may intersect additional tissue in a real patient or in the testing image. In these cases, the calculated attenuation value is possibly incorrect. These are the only rays that are considered when calculating the projection match term.

An example projection is shown in figure 5.9. An image was selected and assigned an arbitrary, exaggerated rigid transformation to simulate the preliminary rigid alignment that occurs before the non-rigid parameters of the model are determined. Applying a rigid transformation to an image before projection is the same as applying the inverse of the transformation to the imaging geometry. This implementation was chosen because it avoids resampling and maintains all the original information about image extents. The region between the two red lines is the region of valid pixels. These are the only pixels that are considered when calculating the projection match function. At the bottom of the image, there is a decrease in intensity. In this region, rays are entering the image volume from the inferior image extent, whereas in an actual patient, they rays would have already passed through some amount of tissue. At the top of the image, the same phenomenon occurs over smaller area. Finally, at the very top of the image, the intensity is zero. Here, the rays missed the source image volume entirely. This projection also shows truncation of the anatomy, where portions of the patient's anatomy are not sampled by any rays. However, this is a geometric limitation of the NST device rather than an issue with this method.

In practice, this is not too large an issue. If the patient is properly positioned, the PBR will be near to the isocenter of the imaging system and the PBR will be sampled with valid rays in sufficiently many projections.

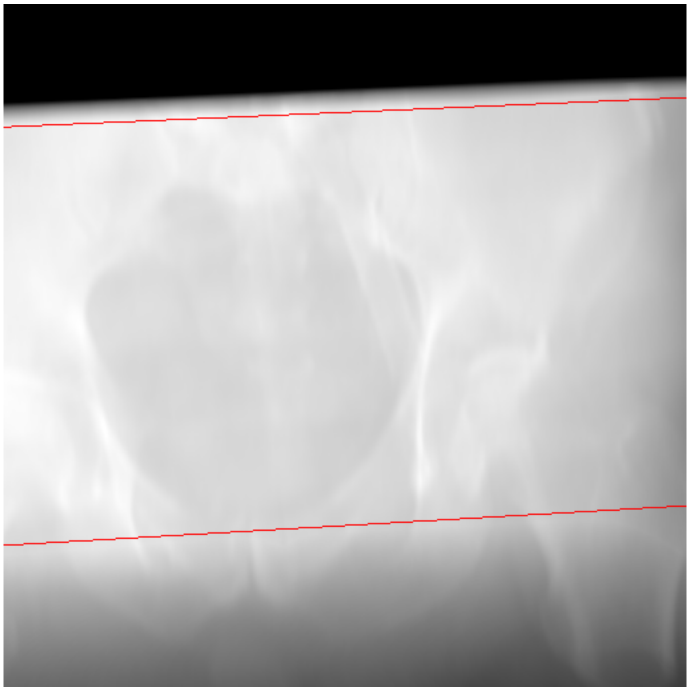


Figure 5.9: Example projection image showing issues associated with truncated source images. The red lines indicate the boundary of detector bins that are valid for use. The regions of decreasing intensity show rays that passed through some tissue in the source image but then exited the volume. The dark region at the top indicates rays that did not intersect the source volume at all.

5.4 Results

In this section, the performance of the method in this work is evaluated, along with that of several variants. Firstly, a short summary of several variants of the method and their intended contribution is provided. Secondly, the common context in which all the method variants are evaluated is discussed. Finally, results are provided for each of the method variants. After this, an additional experiment is described which intends to determine the number of training images that are required to develop a model with sufficient accuracy.

5.4.1 Method Variants

The cumulative method that was developed in this work consists of several parts each of which increase the ability of the method to segment the PBR organs from limited angle images. The first variant explored is the method proposed in this work. This method consists of rigid transformation, a poly-rigid transformation, a skin transformation, a PBR transformation, and a residual transformation, with each of the transformations determined by methods discussed in this chapter. This model is then used to segment the PBR from limited angle images. It will be shown that this variant has the highest accuracy.

After the proposed method is evaluated, a majority of the contributions in this work are stripped away to provide an evaluation of a baseline, non-rigid, PCA-based deformation model in the male pelvis from a reasonable amount of clinical data. In this variant, a rigid transformation is used to align a non-rigid deformation model to simulated 2D projection images. The parameters of this non-rigid deformation model are then determined, and the resulting . This evaluation demonstrates the necessity of the contributions in this work to provide sufficiently accurate non-rigid deformation models. It will be shown that this variant is often less accurate that using a rigid transformation alone.

Finally, the effects of the poly-rigid model are considered in the context of a poly-rigid and non-rigid deformation model. In chapter 4, the poly-rigid method in this work was evaluated in isolation, and evidence and assertions were provided that using a poly-rigid transformation increases the accuracy of an associated non-rigid deformation model by reducing the amount of variation that needs to be explained by a non-rigid deformation model and by increasing the accuracy of the deformations from which a non-rigid deformation model is learned. Later in this section, this increase in accuracy is evaluated in a more clinical context by using a poly-rigid and non-rigid model to segment the PBR. It will be shown that this method produces segmentations that are more accurate than the method described in the previous paragraph and could potentially be considered clinically useful, but this method is not as accurate as the method suggested by this work.

5.4.2 Context of Results

Because the primary goal of this work is to obtain segmentations of the PBR, organ overlap in the PBR organs as quantified by Dice's similarity coefficient (DSC) is used as the evaluation criterion. DSC is defined as

$$\mathrm{DSC}\left(A, B\right) = \frac{2\left|A \cap B\right|}{\left|A\right| + \left|B\right|} \tag{5.4}$$

That is, the volume of intersection of the ground truth manual segmentation and the recovered segmentation divided by the average of volumes of the two segmentations.

The performance of each variant of the method is compared against a benchmark segmentation obtained by rigidly transferring the planning segmentation to the treatment-time space based on a rigid registration obtained from 3D/2D registration between the planning image and each of the test images, this being the only other clinically available method for IGRT in the male pelvis from limited angle images. The results for rigid only alignment to which we are comparing against are shown in figure 5.10. That said, electromagnetic marker localization techniques, such as that provided by the Calypso system [42], may provide more accurate prostate segmentations than this baseline method because these implanted fiducials are more strongly correlated with prostate position than the available image data, which considers the entire patient in a 3D/2D registration.

The data sets for this work consist of 3 patients, each with 1 planning CT and 16 daily CTs. The data are used in a leave-one-out fashion. The planning CT and 15 of the 16 daily CTs are used the learn the model. This model is then used to segment the left out image. Since there is no real data available from patients with expertly segmented daily CTs and data from a limited angle device, all projections are simulated.

The limited angle imaging geometry is that of the NST device. The NST device consists of 52 independent X-ray sources in four linear arrays of 13 sources spaced 1.1 cm appart arranged in a square with a 27 cm edge length. The source array is located 72.3 cm from a square flat panel detector with edge length of 40.96 cm. The detector was assumed to have been binned into 512 × 512 pixels. See 2.1.3 and the rendering in figure 5.11 for more information. Because this device is gantry mounted, the angle of the gantry with respect to the patient provides an additional degree of freedom to this geometry. This geometry is tested in two orientations for each of the variants evaluated, the case where the primary imaging direction is in the AP direction and the case where the primary imaging direction is in the LR direction. Figure 5.11 shows the approximate patient positioning for the AP and LR imaging directions in the NST imaging device used in this work. The patient position is shown as an isosurface rendering of the patient's bony anatomy. The projection of this image from the source indicated in red is shown on the detector. In principle, the primary imaging direction

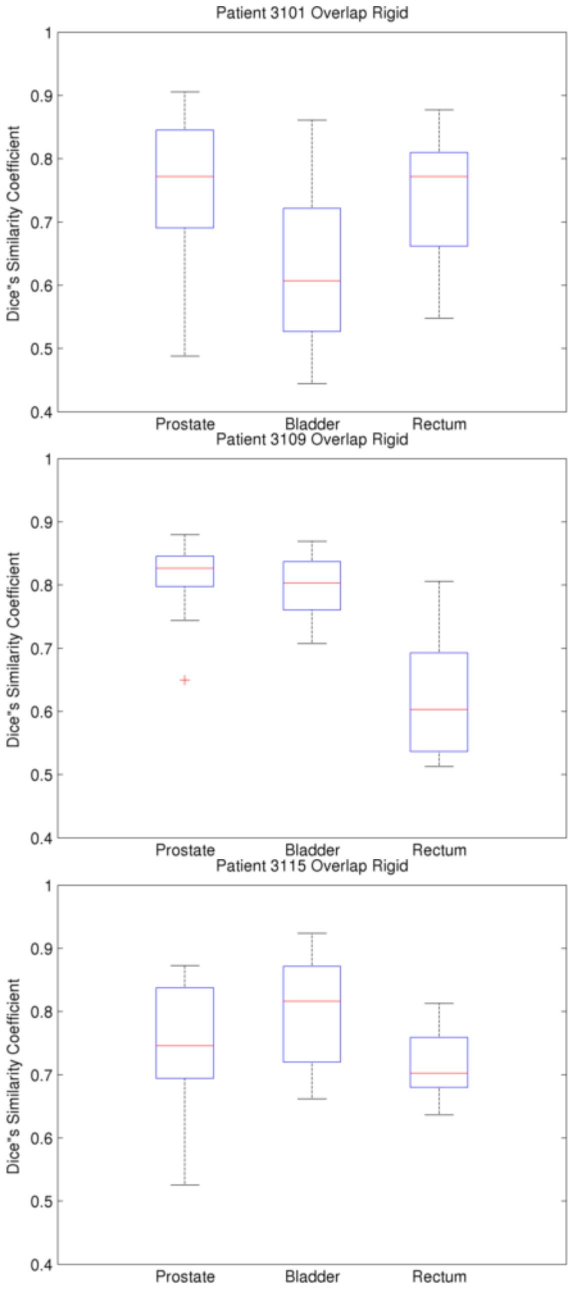
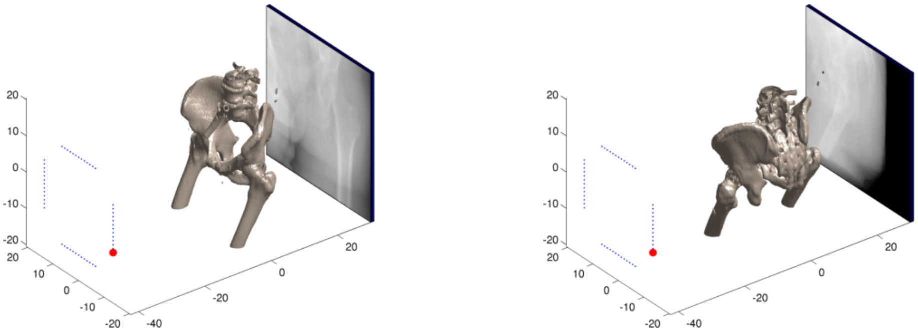


Figure 5.10: DSC between the ground truth segmentation and the planning segmentation rigidly transformed to the treatment space with a transform recovered from 3D/2D registration.



(a) AP Direction

(b) LR Direction

Figure 5.11: Geometry of the limited angle imaging device with approximate patient positioning. The imaging device used is the NST device described in [47] with the gantry positioned such that the predominant imaging direction is in the patient's AP (a) direction and LR direction (b).

could be any direction accessible to the gantry, and there is no reason to assume that AP or LR imaging directions are superior to oblique imaging directions. However, the reason to prefer these orientations has to do with clinical interpretations, as discussed in section 2.2.4. Although this work does not require any image reconstructions, clinical workflow would likely involve reconstructing the limited angle image for review by a clinician. Artifact spread in limited angle images introduces a preferred viewing plane with best resolution that is orthogonal to the primary imaging direction. This preferred viewing plane contains artifacts from anatomy superior and inferior to it. If imaging is performed in the AP or LR direction, the preferred viewing plane is in the coronal or sagittal planes, respectively. Since these correspond to traditional anatomical viewing planes, clinicians will be more familiar the anatomy in these directions than in the oblique directions that would be provided if other orientations are used.

Certain results in the following sections are stated to be significantly improved with respect to the baseline rigid results. This statement is often supported by a pair-wise T-test between the corresponding DSCs. However, DSCs are not Euclidean. That is, improving a DSC from 0.9 to 0.99 is much more significant that improving a DSC from 0.5 to 0.59. Therefore, standard pair-wise T-test is not necessarily the most appropriate hypothesis test. It can be argued that the DSC is a type of probability – the probability that a given point in either segmentation is in both segmentations. The logistic transformation $\log \frac{x}{1-x}$ (synonymously, logit or log-odds) is commonly used to linearize probabilities for linear models. If DSCs are assumed to follow a logit-normal distribution, the standard T-test following a logistic transformation is a more appropriate test for significant difference between paired populations of DSCs. In this work, this logit-normal distribution T-test is used to test for significance. However, both the normal distribution T-test and logit-normal distribution T-test agreed in all instances.

5.4.3 Multi-deformation Model

This section assesses the performance of the complete method developed in this work. Here, a rigid transformation is used along with a poly-rigid deformation model to handle articulated motion of the pelvis and femurs and approximate the deformation that articulation causes to nearby tissue and three deformation models which handle deformation of the skin, PBR, and any residual transformation. The raw data for these models is a set of spatially varying weights for the poly-rigid transformation and a set of transformations from each patient to the atlas image for each of the three non-rigid stages. Dimensionality reduction is then performed on the transformations to develop the deformation models. Here, two variants are assessed, differing only in how the dimensionality reduction is performed. The first variant is the PGA method. The group-wise symmetric Log demons method used in this produces transformations in the Log domain, that is, VVFs. PCA is performed directly on these Log domain transformations. The resulting modes of variation can then be linearly combined, and the resulting Log domain transformation is then Exponentiated to provide a transformation in the typical DVF form. In the PCA method, the Log domain VVFs are first exponentiated prior to performing PCA. The linear modes of variation are then combined to directly produce a DVF.

Placing these two variants on equal footing requires the introduction of an additional element that has the potential to introduce additional error. As described in section 3.4, the models are developed such that the transformations map from the training images to the atlas space. In order to use the models to segment new images, an inverse transformation must be calculated. Recall that the inverse of a Log domain transformation can be computed cheaply and with high precision by negating the VVF. The composite transformation can then be determined by composing each of the inverses in reverse order. This is not the case for the PCA variant, and an explicit inverse must be computed. To address this and ensure that any error from the inverse affects both methods similarly, the inverse method found in [13] is used for both the PGA and PCA cases. This method is simple to implement, fast, and, typically, accurate.

Figure 5.12 shows the results from the PGA variant, and figure 5.13 shows the results from the PCA variant. The mean and median of the DSC values are shown in figure 5.14. Both methods performed well in the AP direction, having mean DSCs comparable to inter-expert variation. Noting that there is greater contrast in MR than in CT, [77] found that the mean inter-expert DSC when segmenting the prostate from preoperative 1.5 T MR images was 0.883 and from intraoperative 0.5 T MR images was 0.838. Based on the available limited angle data, this method likely generates better segmentations than human experts because the prostate is nearly invisible in the reconstructions from the limited angle projection images used for 3D/2D registration. This is shown in figure 5.15.

It was expected that both these variants would perform similarly because the modes of variation that they produce are so similar. However, the PGA variant did not perform

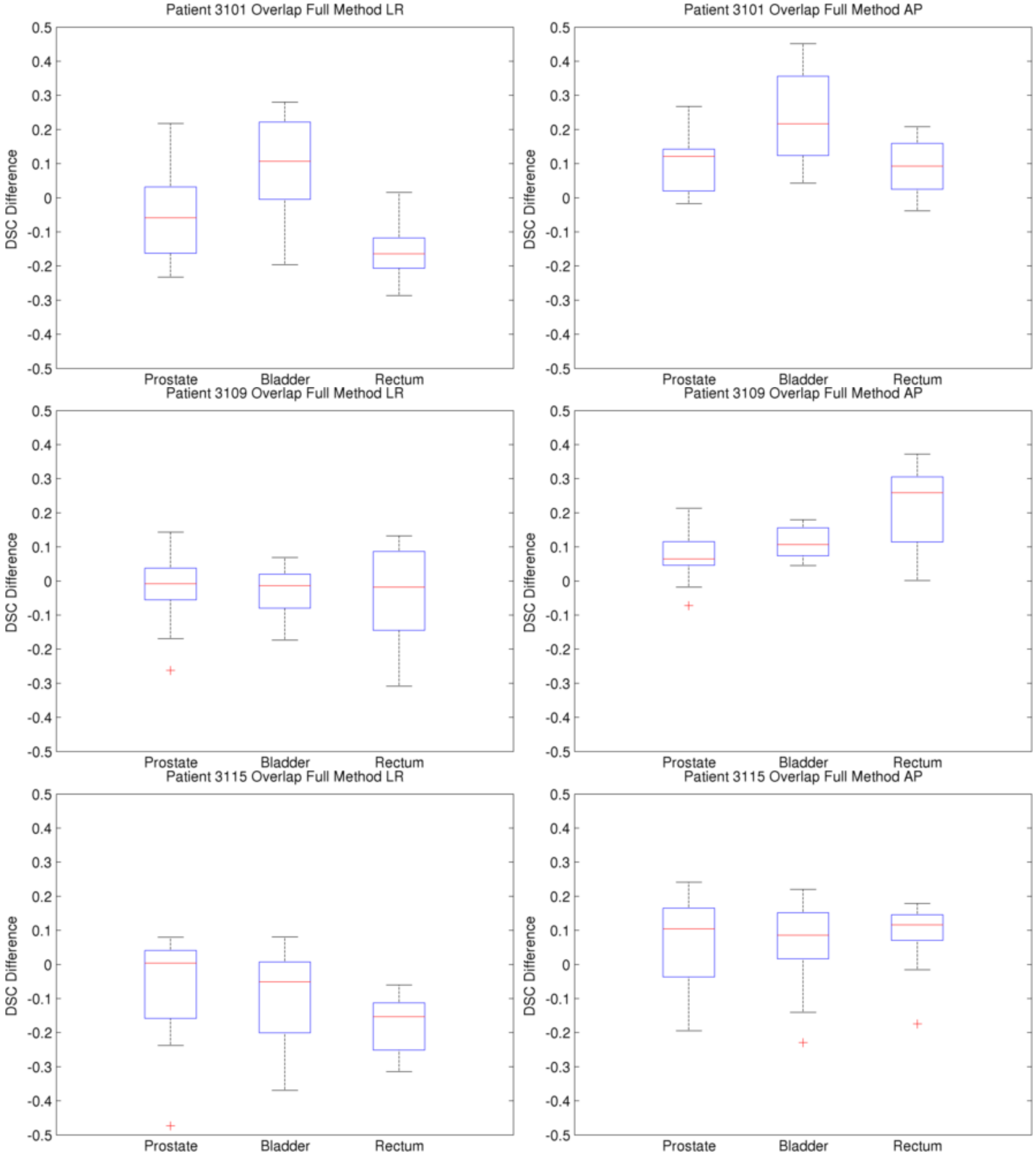


Figure 5.12: Differences in the DSC between the rigid method and the full method proposed in this work. In the AP orientation, the full method provided significant improvement for all organs except the prostate in patient 3115. Performance in the LR orientation was much worse, only providing significant improvement for the bladder in patient 3101.

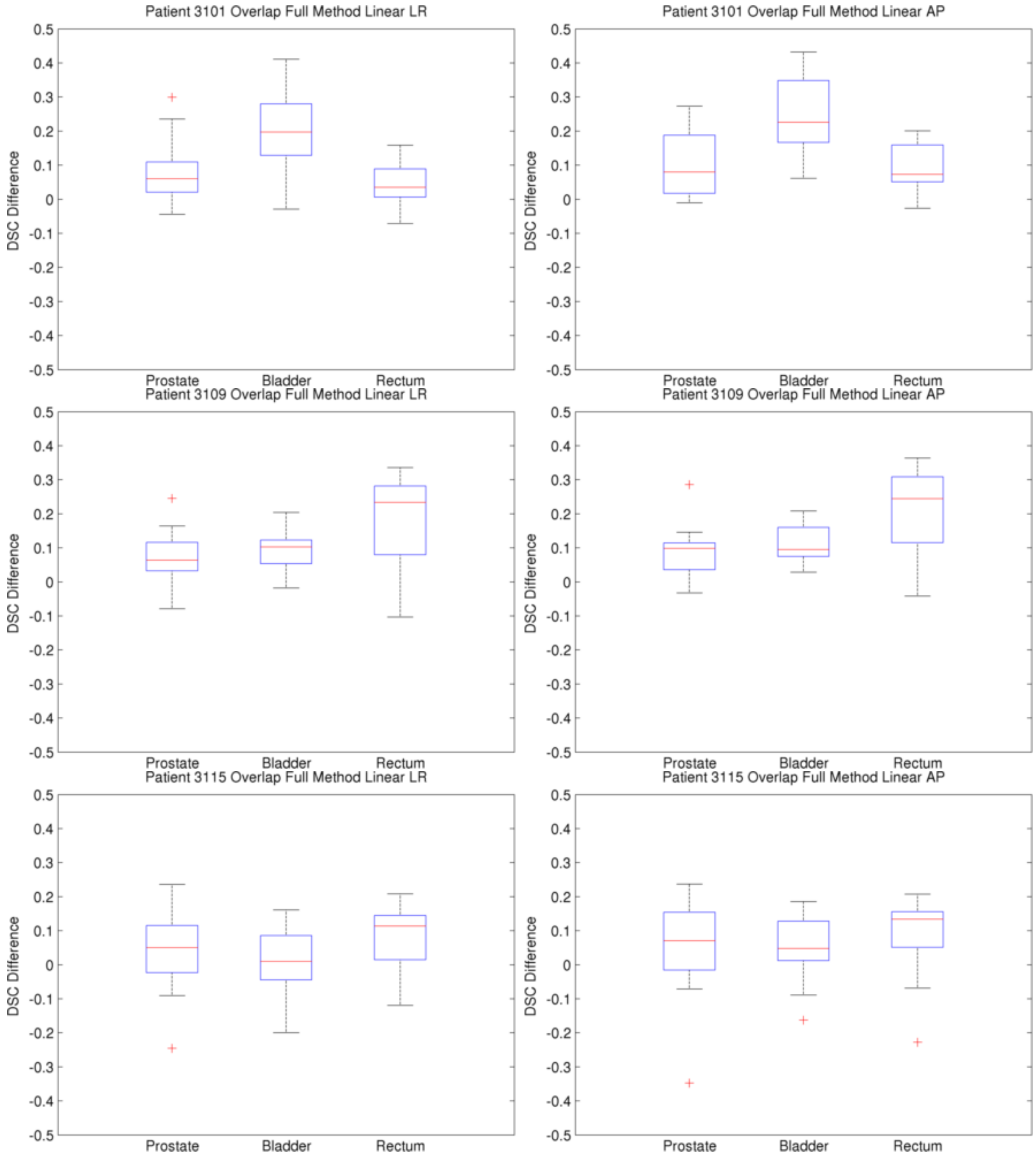


Figure 5.13: Differences in the DSC between the rigid method and the full method proposed in this work using the PCA variants. The method using linear geodesics produced significantly better segmentations than the rigid method for all patients in both orientations except for the prostate in patient 3115 in both orientations and the bladder in patient 3115 in the LR orientations.

	Prostate	Bladder	Rectum
PGA AP	0.851 (0.863)	0.879 (0.905)	0.826 (0.836)
PGA LR	0.726 (0.753)	0.738 (0.763)	0.570 (0.592)
PCA AP	0.850 (0.859)	0.882 (0.907)	0.822 (0.845)
PCA LR	0.835 (0.855)	0.848 (0.870)	0.792 (0.797)

Figure 5.14: Summary of the DSC values (mean with median in parentheses) for each of the organs and both variants and orientations. The variants performed similarly for each of the variants and orientations with the exception of the PGA variant in the LR direction.

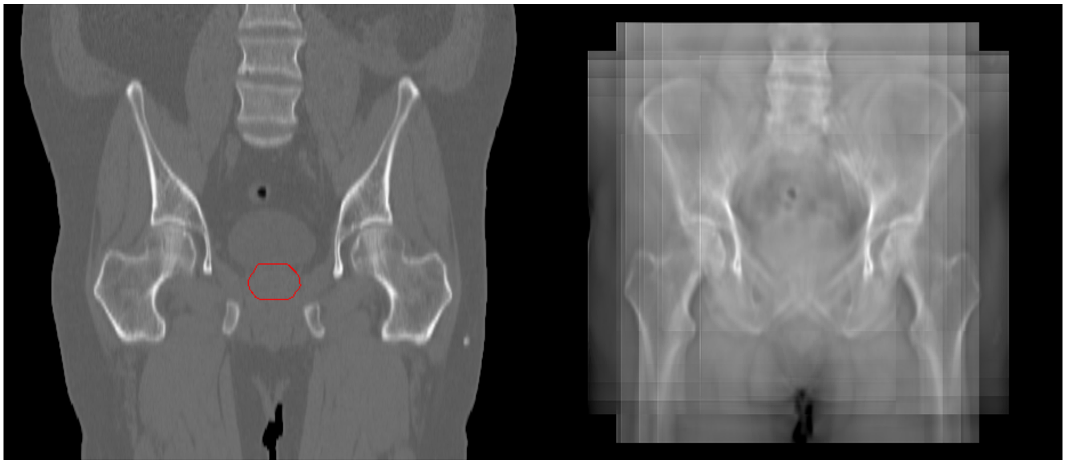


Figure 5.15: Coronal slices of the source CT with prostate indicated in red. The coronal plane being the preferred viewing plane for limited angle images with projections primarily in the AP direction. The CT is then used to simulate limited angle projections for the NST geometry in the AP orientation. These projections are then reconstructed using 10 iterations of the SART reconstruction algorithm. The corresponding slices show the difficulty of identifying the prostate in these limited angle images.

well in the LR direction. Despite their similar performance in the AP orientation and because of this failure, it appears that the use of PGA variant does not provide any advantages over PCA variant in this instance. The parameters of the PCA variant can also be determined faster because it does not require exponentiation during the optimization.

This method is intended to be used as part of an ART protocol. A treatment-time CT (or CT-like image) is necessary to perform dose accumulation for treatment monitoring, which is the central component of ART. However, figure 5.15 demonstrates that limited angle images do not provide reconstructed images of sufficient quality to perform dose accumulation. This method is a deformable segmentation method; it generates a deformation between an atlas image and the treatment space that is used to transfer the atlas segmentation to the treatment space. If this deformation is used to warp the atlas image to the planning space, that warped atlas can serve as an approximation to a CT at treatment-time. The success of this method in segmenting the PBR suggests that the deformed planning image is a good approximation of the patient's actual anatomy because it generates projections that are similar enough to the projections from the actual patient to segment the low contrast PBR structures. Figure 5.16 provides further evidence of this.

5.4.4 Single Deformation Model

This experimental variant evaluates the segmentation performance of a rigid transformation and a single, overall non-rigid deformation model developed using the group-wise symmetric Log demons method presented in this work. This single model attempts to explain all the variation that the patient's pelvic region is likely to undergo during the course of treatment based on the 16 training images. No poly-rigid transformation or additional anatomical information is used. Because the previous data indicated that

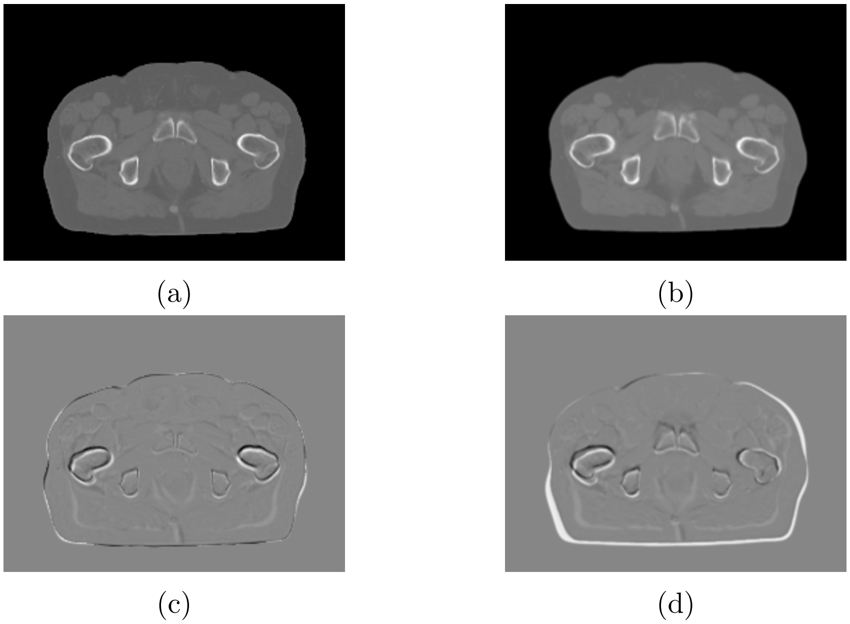


Figure 5.16: a and b show the original patient image warped to the atlas space according the transformation recovered using the method discussed in this section and the rigid transformation recovered from 3D/2D rigid registration, respectively. Both transformations were recovered from projection images taken in the AP orientation. c and d show the difference between the image above them and the atlas image. Observe that errors in the bony anatomy are mostly in the AP direction where the projections provide the least amount of information. The recovered image provides a much better approximation to the patient's anatomy for the purposes of dose calculation than both the reconstructed image shown in figure 5.15 and the image recovered using rigid registration along in d. These results are typical.

the best combination of segmentation accuracy and speed was using PCA in the AP orientation, this variant is only evaluated using those parameters. Figure 5.17 shows the differences between the DSC determined using this simple single deformation model method and the rigid method described previously. The performance of this method is not superior to using rigid transformations only, providing significant improvement only in patient 3109 in the prostate and rectum and often making the segmentation much worse. In terms of accuracy of prostate segmentation, it is better to use rigid transformation alone than to use this variant.

This method corresponds (in that the deformation model is similar to) to several methods [16, 14, 45, 44] that have performed well in the thorax, but it is clearly not successful in the male pelvis. Possible explanations for this have been discussed previously. Basically, deformations in the male pelvis are much more complicated than in the thorax (under respiratory motion). While typically three modes of variation [14] are required to adequately explain the variance in a respiratory motion model, 12 modes of variation were required for the intra-patient data examined here. Since there are only 16 images the training data sets, this indicates that PCA did not produce a particularly general model. It is, however, possible that this variant could succeed if much more data were available.

5.4.5 Poly-rigid and Deformation Model

This next experimental variant uses a poly-rigid transformation along with a single deformation model. The cumulative transformation is then rigid, poly-rigid, and a single non-rigid deformation model. Because the poly-rigid deformation captures articulated motion and approximates the deformation that that articulated motion causes to nearby tissue, these modes of variation no longer need to be explained by the model. This partitions the variance from a single deformation model into two independent

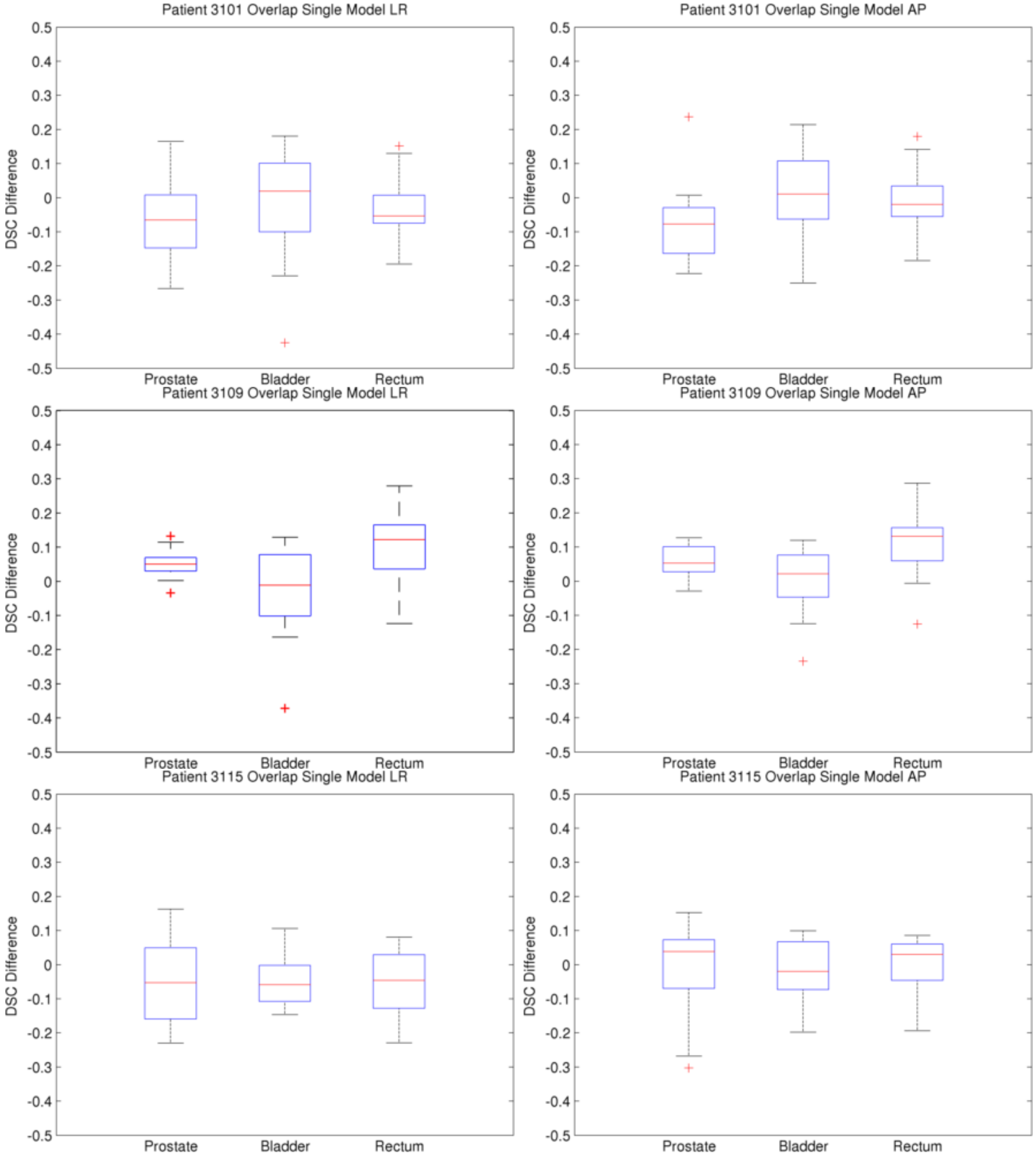


Figure 5.17: Differences in the DSC between the rigid method and a variant of the method in this work using a single deformation model. The single deformation model method only provided significant improvement over the rigid method for patient 3109 in the prostate and rectum for both the AP and LR orientations.

models. As with the previous section, this method is only evaluated with PCA and in the AP orientation. Figure 5.18 shows the differences in DSCs between the rigid baseline and the poly-rigid plus single deformation model. As is expected, with mean DSCs of 0.814, 0.847, and 0.772 for the prostate, bladder, and rectum, respectively, this variant improves on the rigid results of 0.770, 0.743, and 0.693 but does not perform as well as the complete method.

5.4.6 Number of Training Days

In order to implement this method clinically, a patient would need to have daily CTs taken for the first N daily fractions ($N = 15$ based on the data already presented), in addition to the planning image. For the remaining fractions, this method could be used with limited angle images at reduced dose compared to a protocol requiring daily CTs. However, 16 is a significant number of CTs. This section evaluates the performance of the method as the number of daily CTs is decreased from the 16 CTs used in the previous sections. Here, models are learned with an increasing number of training images (from 5 to 13 images, including the planning image), using the models learned from these images to segment the next 4 images in a clinically realistic manner. Figure 5.19 shows the results.

Ultimately, the restriction to 4 test images limits the sensitivity of this analysis. Because there is a limited amount of data available, figure 5.19 is noisy, and no definite conclusions can be drawn from it. The non-increase in performance as a result of increasing the number of training images suggests that more data is needed. The best solution is to increase the number of patients used to test the method, which is necessary for further pre-clinical validation of the method. A combinatorial approach could also be taken where different combinations of training data and test data are evaluated, but this results in more than $100,000$ combinations, making this type of evaluation

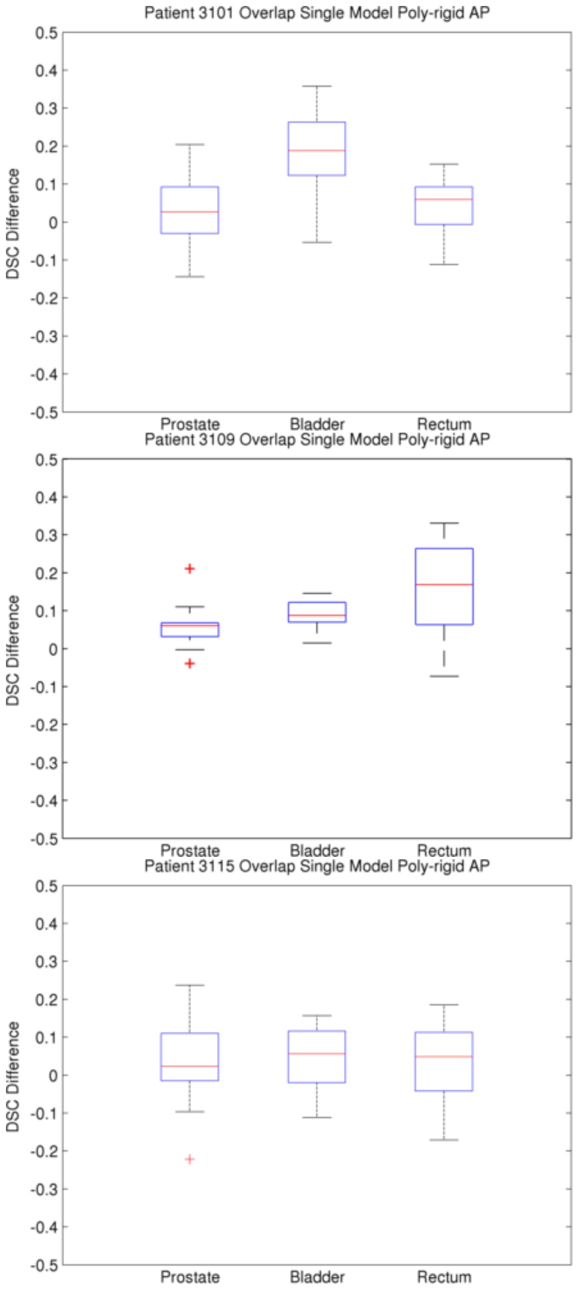


Figure 5.18: DSCs between the ground truth segmentation and the segmentation recovered by the method variant with a poly-rigid transformation and a single non-rigid model.

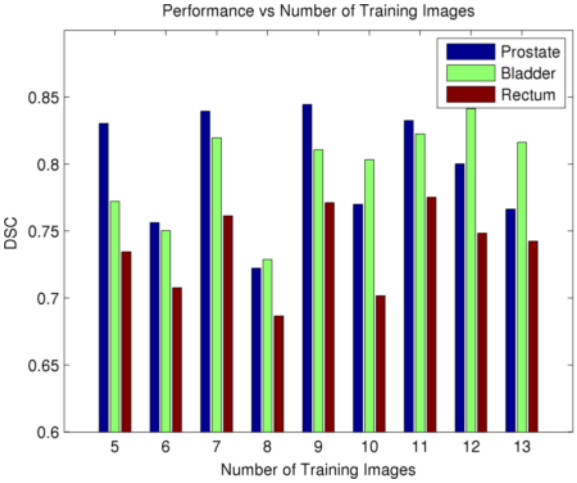


Figure 5.19: This plot shows the absolute performance for the prostate, bladder, and rectum averaged over each of the three patients (for a total of 12 images, four from each patient).

computationally infeasible.

5.5 Summary and Conclusions

Chapter 5 explained the use of multiple non-rigid tissue deformation models to partition variance into several independent models based on anatomical information derived from segmentations. These non-rigid deformation models are assembled together with the poly-rigid transformation explained in chapter 4 to form a complete low dimensionality deformation model that captures sufficient variation to perform intra-subject segmentation of the PBR. It was then explained how to find the parameters of this model from limited angle images. The method was then evaluated in two variants, PGA and PCA, and two orientations, LR and AP. The variants and orientations provided segmentations of the prostate that are comparable with inter-expert segmentation variability, indicating that the method can successfully segment the prostate from limited angles images. It was shown that this method produces a CT-like image, in the form of the deformed atlas image, that is suitably similar to the source CT to perform dose calculation. The calculated dose, together with the deformations produced by the method, can be used for dose accumulation in the atlas space, planning image, or any of the daily images.

Components of the method were then removed to evaluate the contribution of each, producing two variants:

- A fully simplified variant of the method, consisting of a single deformation model and no poly-rigid transformation, was evaluated. This simplified variant did not provide better segmentations than a rigid transformation alone. One is better off using a rigid transformation alone, rather than using a single non-rigid component.

- The poly-rigid transformation explained in chapter 4 was combined with the single deformation model to approximately remove variation due to articulation that would need to be explained by the non-rigid model. As was expected, this variant improved performance over the variant discussed in the previous item but did not perform as well as the full method.

Finally, an attempt was made to determine the number of training CT images that were needed to develop models of sufficient accuracy. The results were inconclusive. Alternative experiments were discussed that may better resolve the question. This work has suggested that 16 images provide models of sufficient accuracy, and it is likely that fewer images than 16 images can be used to develop models that will produce acceptable results. This is modulo the somewhat optimistic results than can be produced by leave-one-out experiments.

Chapter 6

Discussion

6.1 Summary of Contributions

This section addresses the contributions made in chapter 1 and discusses how they were fulfilled by this dissertation.

1. *The development and evaluation of a method that accounts for articulation of rigid structures, and the use of that method to reduce variation to be explained by successive transformations. This method rigidly transforms bony regions and approximates the effects that such bony transformations would have on surrounding tissues.*

   The poly-rigid transformation was discussed in chapter 4. The use of the poly-rigid transformation was motivated by examples showing the failure of linear combinations of rotational transformations to behave as expected and by demonstrations of the ability of the poly-rigid transformation to maintain rigidity in bony anatomy, which is known *a priori* to be true and otherwise not accounted for in PCA or PGA models. Segmentations of the pelvis and femurs in the planning image were used to rigidly segment corresponding bony anatomy in each of the training images. Section 3.3 showed that the Fréchet mean of a poly-rigid transformation is independent of the weight function, which describes the contribution of each transformation to each point in the transformation. Using this knowledge, the mean pose of each of the bones was determined using the Log-Euclidean framework. In this space, a process for selecting a weight function was described that is based on the shapes of the bones in order to approximate the deformation due to articulation in tissues near to the bones. The ability of this method to rigidly align bones was demonstrated by visual comparison of poly-rigidly aligned images versus rigidly aligned images and by a significant decrease in the image distance (SSD) between a template image and rigidly or poly-rigidly aligned images. Furthermore, this improvement in image match continued even after non-rigid registration, indicating that the poly-rigid transformation provides a better initialization for non-rigid registration than rigid transformation alone, likely providing a better non-rigid transformation. Because variation associated with articulated motion has been removed from the training population, that variation no longer needs to be explained by subsequent non-rigid deformation models. The poly-rigid method was evaluated in isolation from the full multiple non-rigid deformation model method in section 5.4.5. There it was demonstrated that using a poly-rigid transformation to remove variation from a single non-rigid deformation model improved the ability of that model to segment the prostate, bladder, and rectum from limited angle images.

2. *The development of an intra-subject deformation model in the male pelvis for use in registration via limited angle imaging that makes use of four models independently accounting variation due to articulation, skin deformation, deformation of the prostate, bladder, and rectum, and residual deformation not accounted for by the previous models. This multi-deformation model paradigm partitions the variance to be explained into several independent models to partially overcome the limitations of PCA.*

   Variation due to articulation was addressed in the previous contribution and chapter 4. Chapter 5 discusses the bulk of this contribution, with introduction and additional information in chapter 3. The construction of the 3D models and their registration to a set of limited angle 2D projection images at treatment-time was discussed. The limitations of PCA were demonstrated by constructing a method based on a rigid transformation and a single deformation model and demonstrating that that method produces inferior results to using rigid transformation alone, in terms of DSCs between ground truth manual segmentations and those found by the method. When the multi-deformation model method that partitions variation into several models was applied, DSCs were significantly increased for all patients, organs, orientations, and PCA/PGA variants, except for the prostate in patient 3109, where rigid transformation alone produced segmentations with DSCs comparable to inter-expert variation (there no improvement could be expected or detected), and the PGA variant in the LR orientation.

3. *A method to improve the convergence of symmetric, group-wise Log demons using the Log-Euclidean Fréchet mean of diffeomorphisms, where the Fréchet mean is a generalization of the Euclidean mean to non-Euclidean spaces.*

   A method for improving the convergence of group-wise registration was developed. This method combines the two approaches for unbiased atlas construction in [38] and [35, 9]: registering the images to an evolving estimate of the mean atlas image and explicitly re-centering the population using methods provided by the Log-Euclidean framework. The method increased convergence in terms of two criteria: the total image distance from the mean atlas image and the total transformation distance from the mean atlas space. This method increased the convergence over either method for every population evaluated. In accordance with the above, this likely indicates better transformations, which are likely to produce better deformation models. The success of this method was demonstrated by the success of the multi-deformation model method at segmenting the PBR and the sharpness of the atlas constructed (a sharper atlas indicating transformations that better match the atlas image).

4. *An evaluation of the usefulness of the Log-Euclidean Framework for dimensionality reduction on diffeomorphisms in the above scenario.*

   The Log-Euclidean framework provides a theoretical advantage over linear treatment of transformations in that group properties are maintained (see section 2.3.2) when treated in the Log domain but are not guaranteed to be maintained when treated in the linear domain. This was demonstration in section 4.1 for rotations and rigid transformations. However, visual observation of the modes of variation indicated a great degree of similarity in the modes of variation, suggesting that the contribution of PGA to statistics on deformations in the context evaluated in this work may be small. This was confirmed by the similarity of results between the PCA and PGA variants evaluated in section 5.4 in the AP direction (although PGA produced insignificantly better results). However, the failure of the PGA variant in the LR direction and the greater speed of the PCA variant suggest that dimensionality reduction on non-rigid transformations may be better considered in the linear domain. Notwithstanding the lack of demonstrated practical benefit from its theoretical advantage in the case of dimensionality reduction of non-rigid transformations, the Log-Euclidean framework provides a great many useful tools instrumental in developing the models used in this work.

5. *An evaluation of the usefulness of the method developed by combining items 1-3 in limited angle registration for IGRT in the male pelvis.*

   The method presented in this work, based on the partitioning of variation into a poly-rigid transformation and three non-rigid deformation models, was evaluated based on its ability to segment the PBR organs from a set of 2D limited angle projection images. Except for the failure cases discussed previously, the method provided significant increases in DSCs for the PBR, with DSCs in the prostate being comparable to inter-expert variation in MR images[77]. The high quality segmentations produced by the method indicate that, With the exception of the speed of the registration at treatment-time, this method is at a stage where it is reasonable to consider further clinical research with the goal of translating the method to clinical practice.

6. *A method for the masking of gas bubbles to increase the accuracy of atlases, registrations, and deformation models learned from registration in regions influenced by transient gas bubbles.*

   The presence of transient gas bubbles in the bowels and, potentially, the bladder causes errors in deformation due to the high contrast bubbles suggesting incorrect correspondence. The types of incorrect deformation indicated was described in section 5.2.4 and decreases the quality of the deformation models. A method was developed to segment the gas bubbles and smoothly interpolate the bubbles with tissue like intensities in such a way that the deformations due to those gas bubbles are captured by the deformation model and as little spurious information in introduced by the masking as possible.

7. *An evaluation of the number of daily CTs required for training data is performed. In a clinical application of this method, a patient would have daily CTs for the first several fractions to provide sufficient data to learn the models. The method could then be used at reduced dose for the remaining fractions.*

   The method was evaluated by using the first N images to learn a model for $N \in [5, 13]$ and then using those models to segment the next four daily images.

> Due to the limited availability of data and computational resources, no conclusions could be drawn on the number of daily images that are required to construct models with sufficient accuracy, except that it appears that models constructed from 16 daily images produce acceptable models (modulo the limitations inherent in leave-one-out studies).

Finally, the thesis statement from chapter 1 is revisited.

Thesis: Intra-patient motion models learned from daily CT images and using multiple, independent statistical deformation models can be used with limited angle projection images to accurately predict the position of the prostate, bladder, and rectum. The correspondence provided by this method allows the estimation of a CT-like image of sufficient quality to enable accumulation of dose from daily fractions into a common coordinate system for treatment monitoring and patient setup adjustment.

It has been demonstrated that the method developed in this work produces intra-patient deformable segmentation that both significantly increases the DSCs in the PBR over rigid only segmentations and produces prostate DSCs that are comparable to inter-expert variation. Furthermore, this method likely produces superior segmentations than could be accomplished by human expert segmenters from the available image data, the PBR being nearly invisible in the images reconstructed from the set of 2D limited angle projections. The success of this method in segmenting the low contrast pelvic organs from 2D projection images at all suggests that the transformed atlas image provides a good surrogate for a treatment-time CT for the purposes of dose calculation. In order for the method to accurately segment the low contrast PBR structures, the method must have produced a deformed image that matches the attenuation in regions surrounding the low contrast structures otherwise their contribution to the 2D projection images would be masked by the contribution of poorly matched attenuation values. This is further supported by visual comparison of an actual CT used for the test and the recovered transformed atlas image in section 5.4.3. The transformations produced by the deformation model during learning and treatment-time can then be used to accumulate the calculated dose to the atlas space or the space of any of the learning or daily images. This fully satisfies the goals of ART from a planning image, a set of learning images, and limited angle daily images at reduced dose and with less manual operation than a protocol based on daily CT images.

6.2 Future Work

This work developed a method for reducing the amount of non-therapeutic imaging dose required to provide the segmentations and CT-like images required for ART by developing a model from a planning CT and several daily CTs for learning. By using limited angle imaging rather than in-room FBCT or device mounted CBCT, imaging time and hardware costs can be decreased, making ART more economically feasible. This section suggests future work regarding improvements to and applications of the method in two sections, method improvements and clinical translation.

6.2.1 Method Improvements

Increasing Accuracy

There is a large body of work over many years focusing on non-rigid registration. Despite this fact, non-rigid registration remains a hard problem. As such, there is a great amount of room for improvement in non-rigid registration methods. An approach used in this work was to improve registration by providing better initialization (in terms of a poly-rigid registration and the better initialization provide to successive registrations by previous registrations) through the introduction of additional information (segmentations). This process of partitioning variance into separate models can be continued and more information can be introduced to potentially increase accuracy.

Different tissues have different mechanical properties and biological function. This reflects the amount and type of deformation that they are likely to undergo. This work has already demonstrated a method for handling the rigid motion undertaken by bones. However, other types of constraints that are more complicated than rigid motion and would need to be handled in the non-rigid stages. The idea is these varied properties can be enforced or encouraged during the regularization stage. For example, the prostate itself is relatively stiff. It does not change shape or volume much, when compared with the bladder and rectum, which greatly change shape and volume. By incorporating this stiffness, the transformations to be learned from are changed for potentially simpler, more specific deformation models. Similarly, fat loss is typically observed during the course of radiotherapy. Therefore, fat regions should shrink more preferentially than, for example muscle or solid organ regions.

There is the additional problem that the amount of fat lost (during a treatment where this is observed) is a function of the time since the initial model construction. Because of this the model may not adequately suggest the actual amount and deformation mode of fat loss. Since fatty tissue can be segmented in the training images, priors could be placed on its decrease in volume. This raises a similar issue in the generalization of this method to other treatment sites. For example, in the head and neck there may be visible tumors that are expected to shrink during treatment. A poly-similarity (or poly-scaling) model, similar to the poly-rigid method used in this work, could be developed to take this fact into account and smoothly approximate the effects of tumor shrinkage on surrounding tissue.

Inter-patient models

The method developed in this work relies on obtaining several daily CTs during their first few fractions to provide training data for learning the intra-patient models. However, the clinically ideal scenario is to acquire a single planning image and use only that image to determine an intra-patient motion model. There are two alternative approaches to solving this problem. This first is to generalize this method to an inter-patient model. As has been demonstrated, there is a significant amount of variation within a single patient. That coupled with the large amount of anatomical variation within humans means that the amount of variation that an inter-patient model would need to explain would be immense. Such a model would also be less specific than an intra-patient model because it would contain parameters that reflect transformations between patients rather than deformations that the target patient is likely to undergo. The space of transformations that patient A can undergo given that he is patient A is much smaller than the space of transformations that patient A can undergo given that he is a member of the much larger class of patients undergoing radiotherapy for prostate cancer. This intra-patient context represents a significant amount of information that is not to be discarded, particularly in the limited angle context, where it is unclear exactly how much freedom can be provided by the model without sacrificing the quality of the results.

A better solution is most likely to develop an intra-patient deformation model from the intra-patient deformation models of other patients. In a general qualitative sense, the types of deformation that a patient undergoes are similar. The bladder fills and empties; the rectum is distended and deforms according to its contents; both move the prostate and surrounding tissues. These types of deformations should be applicable to other patients. A possible assumption that can be made is that patients that are *anatomically similar* will deform similarly, where anatomical similarity is, for example, some combination of the magnitude of deformation between two patients (i.e., the planning image and the atlas image a patient with a training model), perhaps combined with other information. This anatomical similarity could also be computed locally. That is, regions of patients that are anatomically similar to those of the training patients deform similarly. The models from several training patients could then be combined in a weighted fashion to find a model for the new patient.

The major challenge to this task is that there will not be a deformation model that is exactly suited for the novel patient. All the training models must be deformed somehow to match the novel patient. In deforming the atlas space, the deformation models are somehow changed, and determining the nature of these changes remains an open problem.

6.2.2 Clinical Translation

This method produces segmentations of the prostate that are comparable to inter-expert variability from limited angle image data of such limited angular coverage that the organs of interest cannot be identified by a human observer in the reconstructions from this data. This method allows for a dose reduction in a standard ART protocol due to the use of limited angle imaging data for fractions after a sufficient number of training images have been acquired. This method also provides segmentations and CT-like images at reduced manual intervention than a traditional all CT protocol. The success of this method and the data that it provides suggest that research should progress to directing the method to clinical practice. Two major limitations of this method remain and are discussed in the following paragraph.

The first limitation of the method is that the experiments conducted in this work did not lead to a conclusion on the number of the training images that are required to develop models with sufficient accuracy. Additional experiments have been discussed that could answer this question. However, as part of clinical translation this method would be applied to more patients. The results from this application will provide more information about the number of training images that are required. The second limitation is the speed of the method at treatment-time. Despite the fact that the method was developed to use GPUs, which are much faster than central processing units (CPUs) for the types of computation that this method requires (see appendix A), treatment-time application still requires 40-60 minutes of computation. As a first solution to the problem, the computational power of GPUs is increasing rapidly. The current state of the art professional GPU is at least 8× faster than the device used in the experiments conducted for this dissertation. This fact alone makes the method nearly fast enough to be clinically realizable. Because the finite difference and Nelder-Mead optimization approaches used, the function evaluations required for optimization can be scheduled such in order to take advantage of multiple GPUs in the same machine, easily allowing treatment-time computation to be performed in 3-4 minutes. Finally, the code was written in a research context, where clarity, simplicity, and extensibility is preferred over speed, suggesting that there is further room for speed improvements.

APPENDIX A

GPU Acceleration in Medical Imaging

A.1 Introduction

Speed has always been a major concern in any application of computing, both in terms of the number of operations performed and the speed at which those operations can be carried out. Without high-speed computation, the methods developed here can have no clinical value. In this radiotherapy application, a patient must be positioned and remain still during both computation and therapy, occupying both a treatment device and the attention of staff at a very high financial cost. Additionally, if the time of computation following imaging is large, the likelihood of both voluntary and involuntary intra-fractional patient motion is increased. This decreases the utility of the results in clinical application. Furthermore, availability of additional computational power decreases the number of approximating shortcuts that may be made by the developer. This can increase the accuracy of the results. By decreasing the amount of time spent computing online, both the quality and affordability of the treatment is increased. A secondary benefit of speed increase occurs during development. Algorithm development encompasses many stages of construction and test. Although the actual construction (code writing) may be moderately increased when taking advantage of GPU computing, the decrease in time for testing and execution may be great.

A.2 Parallel GPUs in Medical Image Computing

Currently, the number of transistors that can be fit in a given area of an integrated circuit continues to grow exponentially with a doubling time of approximately 2 years

— Moore's Law. When originally stated the number of transistors in the circuit was strongly correlated with performance. The limit of that equivalence has already been reached. CPU designers have traditionally dealt with this problem by increasing the complexity and clock frequency of their processors, but it has become very difficult to build ever larger single CPUs with ever increasing performance. As a result, hardware designers have begun placing multiple cores, copies of the same CPU on the same die. In order to take advantage of multi-core architectures, developers must break code down into separate tasks that can be executed simultaneously. This can be challenging for certain applications.

GPU manufacturers have also embraced this multi-core principle. However, instead of designing complicated processors, they have opted to design simpler processors. These sacrifice some of the ease of traditional sequential programming and make certain operations, such as scattered memory access, more difficult and time consuming, but the amount of die area saved by this simplification can be significant. This die area can be rededicated where it is needed most for numerical computing – floating point mathematics. Furthermore, GPUs take advantage of data parallelism, where the same set of instructions are carried out on the different data. This allows cores to share a large number of die area. As a result of the above, while a modern CPU may have only 8 cores, a modern GPU may have as many as 512. In practice, the performance of certain algorithms can be easily increased by a factor of 300 over their sequential CPU implementation. CPU manufactures dispute this evidence with the argument that, when comparing a heavily optimized, multi-threaded CPU implementation to a heavily optimized GPU implementation, the performance increase is closer to 5-10 times. However, this type of optimization is not usually performed by scientific programmers, but it is apparent in certain libraries. For example, FFTW3 [27], a popular FFT library, is heavily optimized to make use of CPU features. A corresponding GPU library CuFFT shows less benefit over FFTW3 than naive code. As such, heavily FFT-based algorithms will show a much smaller performance gain. That said, many of the algorithms in image analysis and graphics map very well to GPUs and large performance gains can be expected for type of code that is typically written by scientists and engineers.

GPU acceleration is not entirely a holy grail. Some problems are inherently difficult to parallelize. Others do not map well into the hardware available on current GPUs. However, the vast majority of medical image computing is embarrassingly parallel. That is, almost no effort is required to parallelize them. For example, in computing projections, each line integral can be computed independently, in parallel to all the others, since they do not exchange information. All that is required is proper organization of the problem to make best use of the GPU cache. Similarly, in computing a transformations, a displacement vector must be read from memory or evaluated, the corresponding value in the moving image is to be looked up, and then the result is stored. Even problems that are not quite embarrassing, such as reductions where a long vector is reduced to a scalar, as in evaluation of an image match function, have simple, long understood solutions.

In medical image computing, problem sizes are generally large, computation times are great, and results are demanded quickly. The benefits of GPU acceleration here are significant. This can be coupled with the relative ease of porting existing applications to the GPU and the low price of commodity GPU hardware, which is marketed to 3D computer gaming consumers, to suggest that GPU-based medical imaging applications will become incredibly common in the next few years. It is difficult to foresee a future for solely CPU-based applications in both research and industry.

A.3 CUDA Hardware Abstraction

The following is a succinct and necessarily incomplete summary of the CUDA programming abstraction by summarizing the differences between the CUDA and CPU abstractions. It proceeds mostly as a statement of limitation on CUDA with perhaps a single benefit. These limitations may seem formidable, but, with careful execution organization, growing familiarity with the abstraction makes them almost irrelevant for the experienced CUDA developer. Even if you see no way to fit your application entirely within these limits, often the decrease in performance can be minimal. In all but the worst cases, the developer should still see performance increases that would be very difficult to obtain by further optimization of CPU code.

At the time of writing, the most mature general purpose GPU hardware and API combination is CUDA provided by NVidia. It is frequently desirable for programmers to avoid vendor-specific standards in favor of open ones in order to avoid vendor lock-in and potential obsolescence. However, major differences in architectures between the major manufactures, NVidia and ATI, and the relative immaturity of these techniques means that truly hardware independent code is difficult to properly implement. As a result, a decision has been made to adopt CUDA due to their early entrance into this area, availability of tools, number of users, and continued, rapid hardware and software development towards the end of general purpose GPU computing. This is targeted for current devices with compute capability 2.1 or higher. The compute capability 2.1 abstraction decreases the complexity and increases the flexibility of programming over earlier compute capability.

CUDA is an extension to C/C++ and associated API that controls the GPU and describes the code to be executed by the GPU. Here, the GPU, or device, is treated as a co-processor to the CPU, or host, that has independent memory and allows parallel execution. The actual code executed on the device is specified in a kernel. The program should be organized such the number of parallel executions, threads, should be as large as possible, many thousands. A kernel should, as much as possible, follow the same execution path for all threads. A general kernel will establish its location in the execution configuration by reading its thread specific `threadIdx`, read data from global device memory, perform the task at hand, and write the results back to global memory. Threads are organized into 1-, 2-, or 3-D blocks consisting of a varying number of threads from 32 (the minimum for best hardware utilization) to around 1024, limited by the availability of resources. Several/many blocks are launched simultaneously and distributed to the various multi-processors in the GPU, groups of CUDA core execution units. Blocks are organized into 1-, 2- or 3-D grids. This grid, block, thread structure organizes the kernel execution. This allows implicit scalability of code in that 8 blocks executed on a two multi-processor system are executed two-at-a-time while 8 blocks on an four multi-processor system are executed four-at-a-time. There is no inter-thread communication at between blocks, except through global memory, which can require complicated synchronization, but inter-thread communication is performed within blocks using shared memory.

CUDA hardware employs a hierarchical memory structure based on the constructs to which this memory is visible and the speed at which it can be accessed. Global memory is the top-level memory, is visible to all threads, and can be copied to and from the host. This memory has a large latency, several hundred clock cycles, but this latency can be hidden. It should also be accessed in many word, aligned, sequential chunks. Such a coalesced access pattern allows a larger amount of memory to be transferred in a single global memory transaction. Since device global memory to host main memory has high latency and a much lower bandwidth than any memory on the device, device to host and host to device transfers should be minimized. When they are necessary (i.e., for any results), they should take place in fewer large transactions rather than more small ones. Failure to coalesce global memory transactions results in a increase number of transactions, decreasing effective memory bandwidth. Shared memory can be accessed by all threads within a block and is the best option for inter-thread communication. It can be accessed in a single clock cycle, barring any delays needed for synchronization. Shared memory is organized into banks where access is available without delay as long as individual threads access distinct bins or all access the same thread. Understanding bank conflicts is important for optimal performance, but further rules about bank conflicts are not discussed here. Details can be found in [55]. At the thread level, memory is registers available in a single clock cycle. These are only visible with the thread and are allocated from a fixed pool of registers available to a block. If more registers are required than are available, these are stored in local memory. This is stored in a global memory in an always-coalesced pattern but has the same downsides as global memory. It is to be avoided if possible.

The remaining memories have special features. The first is constant memory. It is read-only within a kernel and is optimized for all threads accessing the same location simultaneously. One of the more valuable memories is texture/surface memory. This memory offers hardware limited-precision texture interpolation, range checking, and integer-to-float scaling. Texture memory does not require strict coalescing and is cached for 2D spatial locality so threads within the same warp (defined below) accessing data that is nearby in 2D achieve performance increase. Texture memory must be copied from linear memory into a `cudaArray` structure so that it can be reorganized for use by the texture hardware. Surface memory is similar to texture memory but it can be written to. Finally, in current GPUs, global memory accesses are cached. This is not explicitly visible to the developer but allows a certain amount of uncoalesced-style memory access without excess memory transactions (e.g., approximating image gradients by central difference) without the limitation of texture memory.

Within a kernel, development is very much like C or C++ with few caveats. Within a thread block, all the threads are subdividing into warps. In standard geek punnery, warp is a term from cloth weaving meaning a group of threads [55]. In the current implementation, the warp size is 32 threads. A warp shares some computing hardware but are free to branch. However, when threads within a warp diverge, execution must be serialized. With a highly branching instruction path, this can lead to great performance loss. Different warps may diverge without penalty. It is important to minimize branching to achieve maximum performance, but simultaneous execution within a warp can lead to simpler synchronization for data transfer between threads. Global memory access also occurs at the warp level, and warp switching is used to hide global memory latency. When one warp is waiting for data, another can be executing instructions meaning that block sizes should be an (preferably large) integer multiple of the warp size.

Of greater value is proper thread organization and distribution of work. In order to ensure that thread blocks are of adequate size to hide latency and grids are of adequate size to keep multi-processors busy, many threads should be launched. For image processing, this usually means one thread per pixel or voxel, and in order to ensure that global memory bandwidth is optimized and coalesced, thread accesses to global memory should be aligned and coalesced. In order to be coalesced, memory accesses can only request a value of 1, 2, 4, 8, or 16 bytes aligned to the size of its data type. So, for example, a standard displacement vector field access $[x_0 y_0 z_0 \dots x_{n-1} y_{n-1} z_{n-1}]$ should be arranged as $[x_0 \dots x_{n-1} y_0 \dots y_{n-1} z_0 \dots z_{n-1}]$ for optimal throughput. Further details are mentioned in [55].

The above limitations and the 2D locality of texture memory give rise to two common access patterns. These are provided in CUDA C/C++ both for clarity and to provide some example of CUDA code. Three CUDA specific constructs are demonstrated.

```
void do_something_gpu(float *d_out, float *d_a, float *d_b, int len) {
  //define 1D blocks and grid
  //these should be determined experimentally on a device-by-device
  //basis, but these values work well for large arrays
  unsigned int n_threads = 256; //block size
  unsigned int n_blocks = 64; //grid size

  //launch kernel
  do_something_gpu_kernel<<<n_blocks, n_threads>>>(d_out, d_in, len);
}

__global__ void do_something_gpu_kernel(float *d_out, float *d_in, int
    len) {
  //get position of current thread within block
  int thread_id = threadIdx.x;
  //position of block within grid
  int block_id = blockIdx.x;
  //starting position of current thread
  int position = block_id*blockDim.x+thread_id;
  //amount to increment index variable to next element
  int increment = blockDim.x*gridDim.x;

  while(position<len) {
    //each thread in thread block access adjacent values
    d_out[position] = something(d_in[position]);
    position += increment;
  }
}
__device__ float something(float x) {
  ...
}
```

Source A.1: 1D CUDA access pattern

The first is `<<<n_blocks, n_threads>>>` which specifies the execution configuration. `__global__` and `__device__` specify that the function specifies a kernel and a function for execution by a kernel. The first sample describes an array access pattern which transforms an array of floats according to function a function that is independent of any 3D location. The second divides each xy slice into rectangular blocks and then marches along the z axis. These are probably the most common access patterns and well utilize resources.

```
void do_something_gpu_3d(float *d_out, float *d_a, float *d_b, int dimx,
    int dimy, int dimz) {
  //define 2D blocks and grid
  dim3 n_threads(32,8); //block size
  dim3 n_blocks; //grid size
  n_blocks.x=(dimx%n_threads.x==0)?dimx/n_threads.x:dimx/n_threads.x+1;
  n_blocks.y=(dimy%n_threads.y==0)?dimy/n_threads.y:dimx/n_threads.y+1;
  n_blocks.z=1;

  //launch kernel
  do_something_gpu_kernel_3d<<<n_blocks, n_threads>>>(d_out, d_in, dimx,
     dimy, dimz);
}

__global__ void do_something_gpu_kernel_3d(float *d_out, float *d_in, int
    dimx, int dimy, int dimz) {
  //get position of current thread within block
  int thread_id_x = threadIdx.x;
  int thread_id_y = threadIdx.y;
  //position of block within grid
  int block_id_x = blockIdx.x;
  int block_id_y = blockIdx.y;
  //starting position of current thread
  int x = block_id_x*blockDim.x+thread_id_x;
  int y = block_id_y*blockDim.y+thread_id_y;
  int position = y*dimx+x;

  //amount to increment index variable to next element
  int rowlen = blockDim.x*gridDim.x;
  int collen = blockDim.y*gridDim.y;
  int increment = rowlen*collen;

  //make sure you are in range of the slice
  if(x<dimx && y<dimy) {
    //do successive z slabs
    for(int z = 0; z<dimz; z+=blockSize.z) {
      d_out[position] = something_3d(x,y,z,d_in[position]);
      position += increment;
    }
  }
}

__device__ float something_3d(int x, int y, int z, float v) {
  ...
}
```

Source A.2: 2D/3D CUDA access pattern

BIBLIOGRAPHY

[1] AH Andersen and AC Kak. Simultaneous algebraic reconstruction technique (sart): a superior implementation of the art algorithm. *Ultrasonic imaging*, 6(1):81–94, 1984.

[2] V. Arsigny, O. Commowick, X. Pennec, and N. Ayache. A Log-Euclidean framework for statistics on diffeomorphisms. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006*, pages 924–931, 2006.

[3] V. Arsigny, X. Pennec, and N. Ayache. Polyrigid and polyaffine transformations: A new class of diffeomorphisms for locally rigid or affine registration. *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2003*, pages 829–837, 2003.

[4] V. Arsigny, X. Pennec, and N. Ayache. Polyrigid and polyaffine transformations: a novel geometrical tool to deal with non-rigid deformations-application to the registration of histological slices. *Medical Image Analysis*, 9(6):507–523, 2005.

[5] H.H. Barrett and K.J. Myers. *Foundations of image science*. Wiley, 2004.

[6] Kanwal K Bhatia, Joseph V Hajnal, Basant K Puri, A David Edwards, and Daniel Rueckert. Consistent groupwise non-rigid registration for atlas construction. In *Biomedical Imaging: Nano to Macro, 2004. IEEE International Symposium on*, pages 908–911. IEEE, 2004.

[7] J. Binder and W. Kramer. Robotically-assisted laparoscopic radical prostatectomy. *BJU International*, 87(4):408–410, 2001.

[8] Andreas Blana, Bernhard Walter, Sebastian Rogenhofer, Wolf F Wieland, et al. High-intensity focused ultrasound for the treatment of localized prostate cancer: 5-year experience. *Urology*, 63(2):297, 2004.

[9] M. Bossa, M. Hernandez, and S. Olmos. Contributions to 3D diffeomorphic atlas estimation: Application to brain images. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2007*, pages 667–674, 2007.

[10] M. Bossa and S. Olmos. A new algorithm for the computation of the group logarithm of diffeomorphisms. In *2nd MICCAI Workshop on Mathematical Foundations of Computational Anatomy*, October 2008.

[11] P. Cachier and X. Pennec. 3D non-rigid registration by gradient descent on a Gaussian-windowed similarity measure using convolutions. In *Mathematical Methods in Biomedical Image Analysis, 2000. Proceedings. IEEE Workshop on*, pages 182–189. IEEE, 2000.

[12] Joshua Cates, P Thomas Fletcher, Martin Styner, Martha Shenton, and Ross Whitaker. Shape modeling and analysis with entropy-based particle systems. In *Information Processing in Medical Imaging*, pages 333–345. Springer, 2007.

[13] M. Chen, W. Lu, Q. Chen, K.J. Ruchala, and G.H. Olivera. A simple fixed-point approach to invert a deformation field. *Medical physics*, 35:81, 2008.

[14] C. Chou, B. Frederick, X. Liu, G. Mageras, S. Chang, and S. Pizer. CLARET: A fast deformable registration method applied to lung radiation therapy. In *Proc. Fourth International MICCAI Workshop on Pulmonary Image Analysis (PULMO 2011)*, Toronto, Canada, September 2011.

[15] Chen-Rui Chou and Stephen Pizer. Real-time 2d/3d deformable registration using metric learning. In *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging*, pages 1–10. Springer, 2013.

[16] C.R. Chou, C. Frederick, S. Chang, and S. Pizer. A learning-based patient repositioning method from limited-angle projections. *Brain, Body and Machine*, pages 83–94, 2010.

[17] G.E. Christensen, S.C. Joshi, and M.I. Miller. Volumetric transformation of brain anatomy. *Medical Imaging, IEEE Transactions on*, 16(6):864–877, 1997.

[18] Marc A. Dall'Era, Matthew R. Cooperberg, June M. Chan, Benjamin J. Davies, Peter C. Albertsen, Laurence H. Klotz, Christopher A. Warlick, Lars Holmberg, Donald E. Bailey, Meredith E. Wallace, Philip W. Kantoff, and Peter R. Carroll. Active surveillance for early-stage prostate cancer. *Cancer*, 112(8):1650–1659, 2008.

[19] B.C. Davis, M. Foskey, J. Rosenman, L. Goyal, S. Chang, and S. Joshi. Automatic segmentation of intra-treatment ct images for adaptive radiation therapy of the prostate. In JamesS. Duncan and Guido Gerig, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2005*, volume 3749 of *Lecture Notes in Computer Science*, pages 442–450. Springer Berlin Heidelberg, 2005.

[20] R. Deriche. Fast algorithms for low-level vision. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 12(1):78–87, 1990.

[21] J.T. Dobbins III and D.J. Godfrey. Digital x-ray tomosynthesis: current state of the art and clinical potential. *Physics in medicine and biology*, 48(19):R65, 2003.

[22] P. Felzenszwalb and D. Huttenlocher. Distance transforms of sampled functions. Technical report, Cornell University, 2004.

[23] Frederico Ferronha, Fortunato Barros, Victor Vaz Santos, Vincent Ravery, and Vincent Delmas. Is there any evidence of superiority between retropubic, laparoscopic or robot-assisted radical prostatectomy? *International braz j urol*, 37(2):146–160, 2011.

[24] P.T. Fletcher, C. Lu, S.M. Pizer, and S. Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *Medical Imaging, IEEE Transactions on*, 23(8):995–1005, 2004.

[25] R Fletcher. *Practical Methods of Optimization*. Wiley, 1987.

[26] B. Frederick, D. Lalush, and S. Chang. Registration using nanotube stationary tomosynthesis: Comparison of 3D/3D to 3D/2D methods. *Medical Physics*, 37:3460, 2010.

[27] M. Frigo and S.G. Johnson. The design and implementation of FFTW3. *Proceedings of the IEEE*, 93(2):216–231, 2005.

[28] Guido Gerig, Martin Styner, D Jones, Daniel Weinberger, and Jeffrey Lieberman. Shape analysis of brain ventricles using spharm. In *Mathematical Methods in Biomedical Image Analysis, 2001. MMBIA 2001. IEEE Workshop on*, pages 171–178. IEEE, 2001.

[29] D.J. Godfrey, F.F. Yin, M. Oldham, S. Yoo, and C. Willett. Digital tomosynthesis with an on-board kilovoltage imaging device. *International Journal of Radiation Oncology, Biology, and Physics*, 65(1):8–15, 2006.

[30] D.G. Grant. Tomosynthesis: A three-dimensional radiographic imaging technique. *Biomedical Engineering, IEEE Transactions on*, 19(1):20–28, 1972.

[31] X. Gu, D. Choi, C. Men, H. Pan, A. Majumdar, and S.B. Jiang. Gpu-based ultrafast dose calculation using a finite size pencil beam model. *Physics in medicine and biology*, 54(20):6287, 2009.

[32] Peter Hammerer and Stephan Madersbacher. Landmarks in hormonal therapy for prostate cancer. *BJU International*, 110:23–29, 2012.

[33] A.C. Hartford, J.M. Galvin, D.C. Beyer, T.J. Eichler, G.S. Ibbott, B. Kavanagh, C.J. Schultz, and S.A. Rosenthal. American college of radiology (ACR) and american society for radiation oncology (astro) practice guideline for intensity-modulated radiation therapy (IMRT). *American journal of clinical oncology*, 35(6):612–617, 2012.

[34] D. Hayne, CJ Vaizey, and PB Boulos. Anorectal injury following pelvic radiotherapy. *British journal of surgery*, 88(8):1037–1048, 2001.

[35] Monica Hernandez, Matias Bossa, and Salvador Olmos. Estimation of statistical atlases using groups of diffeomorphisms. Technical report, Technical report, I3A, University of Zaragoza, 2007.

[36] Lars Holmberg, Anna Bill-Axelson, Fred Helgesen, Jaakko O. Salo, Per Folmerz, Michael Haggman, Swen-Olof Andersson, Anders Spangberg, Christer Busch, Steg

Nordling, Juni Palmgren, Hans-Olov Adami, Jan-Erik Johansson, and Bo Johan Norlen. A randomized trial comparing radical prostatectomy with watchful waiting in early prostate cancer. *New England Journal of Medicine*, 347(11):781–789, 2002. PMID: 12226148.

[37] D.A. Jaffray, J.H. Siewerdsen, J.W. Wong, and A.A. Martinez. Flat-panel cone-beam computed tomography for image-guided radiation therapy. *International Journal of Radiation Oncology* Biology* Physics*, 53(5):1337–1349, 2002.

[38] S Joshi, Brad Davis, Matthieu Jomier, and Guido Gerig. Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage*, 23:S151–S160, 2004.

[39] A.C. Kak and M. Slaney. *Principles of computerized tomographic imaging.* IEEE Service Center, Piscataway, NJ, 1988.

[40] F.M. Khan. *The physics of radiation therapy.* Lippincott Williams & Wilkins Philadelphia, 4th edition, 2010.

[41] D. Knaan and L. Joskowicz. Effective intensity-based 2D/3D rigid registration between fluoroscopic X-ray and CT. *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2003*, pages 351–358, 2003.

[42] Patrick Kupelian, Twyla Willoughby, Arul Mahadevan, Toufik Djemil, Geoffrey Weinstein, Shirish Jani, Charles Enke, Timothy Solberg, Nicholas Flores, David Liu, et al. Multi-institutional clinical experience with the calypso system in localization and continuous, real-time monitoring of the prostate gland during external radiotherapy. *International Journal of Radiation Oncology* Biology* Physics*, 67(4):1088–1098, 2007.

[43] Joshua H Levy, Mark Foskey, and Stephen M Pizer. Rotational flows for interpolation between sampled surfaces. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–8. IEEE, 2008.

[44] R. Li, X. Jia, J.H. Lewis, X. Gu, M. Folkerts, C. Men, and S.B. Jiang. Real-time volumetric image reconstruction and 3d tumor localization based on a single x-ray projection image for lung cancer radiotherapy. *Medical Physics*, 37(6):2822–2826, 2010.

[45] X. Liu, B.C. Davis, M. Niethammer, S.M. Pizer, and G.S. Mageras. Prediction-driven respiratory motion atlas formation for 4d image-guided radiation therapy in lung. In *MICCAI, International Workshop on Pulmonary Image Analysis*, 2010.

[46] X. Liu, R.R. Saboo, S.M. Pizer, and G.S. Mageras. A shape-navigated image deformation model for 4D lung respiratory motion estimation. In *Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on*, pages 875–878. IEEE, 2009.

[47] J.S. Maltz, F. Sprenger, J. Fuerst, A. Paidi, F. Fadler, and A.R. Bani-Hashemi. Fixed gantry tomosynthesis system for radiation therapy image guidance based on a multiple source x-ray tube with carbon nanotube cathodes. *Medical physics*, 36:1624, 2009.

[48] A.A. Martinez, D. Yan, D. Lockman, D. Brabbins, K. Kota, M. Sharpe, D.A. Jaffray, F. Vicini, and J. Wong. Improvement in dose escalation using the process of adaptive radiotherapy combined with three-dimensional conformal or intensity-modulated beams for prostate cancer. *International Journal of Radiation Oncology* Biology* Physics*, 50(5):1226–1234, 2001.

[49] K I M McKinnon. Convergence of the nelder–mead simplex method to a nonstationary point. *SIAM Journal on Optimization*, 9(1):148–158, 1998.

[50] Virginia A. Moyer. Screening for prostate cancer: U.s. preventive services task force recommendation statement. *Annals of Internal Medicine*, 157(2):120–134, 2012.

[51] M.J. Murphy, J. Balter, S. Balter, J.A. BenComo Jr, I.J. Das, S.B. Jiang, C.M. Ma, G.H. Olivera, R.F. Rodebaugh, K.J. Ruchala, et al. The management of imaging dose during image-guided radiotherapy: Report of the AAPM task group 75. *Medical physics*, 34:4041, 2007.

[52] R.M. Murray, Z. Li, and S.S. Sastry. *A mathematical introduction to robotic manipulation.* CRC, 1994.

[53] John A Nelder and Roger Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

[54] J. Nijkamp, F.J. Pos, T.T. Nuver, R. de Jong, P. Remeijer, J.J. Sonke, and J.V. Lebesque. Adaptive radiotherapy for prostate cancer using kilovoltage cone-beam computed tomography: first clinical results. *International Journal of Radiation Oncology* Biology* Physics*, 70(1):75–82, 2008.

[55] NVidia. *NVIDIA CUDA Programming Guide*, volume Version 4.0. NVidia, 2011.

[56] GM Onik, JK Cohen, GD Reyes, B Rubinsky, Z Chang, and J Baust. Transrectal ultrasound-guided percutaneous radical cryosurgical ablation of the prostate. *Cancer*, 72(4):1291, 1993.

[57] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, SIGGRAPH '03, pages 313–318, New York, NY, USA, 2003. ACM.

[58] S.M. Pizer, P.T. Fletcher, S. Joshi, A.G. Gash, J. Stough, A. Thall, G. Tracton, and E.L. Chaney. A method and software for segmentation of anatomic object ensembles by deformable m-reps. *Medical Physics*, 32:1335, 2005.

[59] S.P. Poplack, T.D. Tosteson, C.A. Kogel, and H.M. Nagy. Digital breast tomosynthesis: initial experience in 98 women with abnormal digital screening mammography. *American Journal of Roentgenology*, 189(3):616–623, 2007.

[60] L. Ren, D.J. Godfrey, H. Yan, Q.J. Wu, and F.F. Yin. Automatic registration between reference and on-board digital tomosynthesis images for positioning verification. *Medical physics*, 35:664, 2008.

[61] L. Ren, J. Zhang, D. Thongphiew, D.J. Godfrey, Q.J. Wu, S.M. Zhou, and F.F. Yin. A novel digital tomosynthesis (DTS) reconstruction method using a deformation field map. *Medical physics*, 35:3110, 2008.

[62] Daniel Rueckert, Luke I Sonoda, Carmes Hayes, Derek L. G. Hill, Martin O. Leach, and David J. Hawkes. Nonrigid registration using free-form deformations: application to breast mr images. *Medical Imaging, IEEE Transactions on*, 18(8):712–721, 1999.

[63] Rohit Saboo. *Atlas Diffeomorphisms via Object Models*. PhD thesis, University of North Carolina at Chapel Hill, 2011.

[64] Kaleem Siddiqi and Stephen M Pizer. *Medial representations: mathematics, algorithms and applications*, volume 37. Springer, 2008.

[65] R. Siegel, E. Ward, O. Brawley, and A. Jemal. Cancer statistics, 2011. *CA: a cancer journal for clinicians*, 2011.

[66] American Cancer Society. Expectant management (watchful waiting) and active surveillance for prostate cancer. http://www.cancer.org/cancer/prostatecancer/detailedguide/prostate-cancer-treating-watchful-waiting. Accessed: 02/13/2013.

[67] Ian F Tannock, David Osoba, Martin R Stockler, D Scott Ernst, Alan J Neville, Malcolm J Moore, George R Armitage, Jonathan J Wilson, Peter M Venner, CM Coppin, et al. Chemotherapy with mitoxantrone plus prednisone or prednisone alone for symptomatic hormone-resistant prostate cancer: a canadian randomized trial with palliative end points. *Journal of Clinical Oncology*, 14(6):1756–1764, 1996.

[68] H.K. Tuy. An inversion formula for cone-beam reconstruction. *SIAM Journal on Applied Mathematics*, pages 546–552, 1983.

[69] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache. Non-parametric diffeomorphic image registration with the demons algorithm. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2007*, pages 319–326, 2007.

[70] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache. Symmetric Log-domain diffeomorphic registration: A demons-based approach. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2008*, pages 754–761, 2008.

[71] J. Vikgren, S. Zachrisson, A. Svalkvist, Å.A. Johnsson, M. Boijsen, A. Flinck, S. Kheddache, and M. Båth. Comparison of chest tomosynthesis and chest radiography for detection of pulmonary nodules: Human observer study of clinical cases. *Radiology*, 249(3):1034–1041, 2008.

[72] Patrick C Walsh, Penny Marschke, Deborah Ricker, and Arthur L Burnett. Patient-reported urinary continence and sexual function after anatomic radical prostatectomy. *Urology*, 55(1):58–61, 2000.

[73] W. Wein, B. Röper, and N. Navab. 2D/3D registration based on volume gradients. In *SPIE Medical Imaging*, volume 5747, pages 144–150, 2005.

[74] Q.J. Wu, D. Thongphiew, Z. Wang, B. Mathayomchan, V. Chankong, S. Yoo, W.R. Lee, and F.F. Yin. On-line re-optimization of prostate IMRT plans for adaptive radiation therapy. *Physics in medicine and biology*, 53(3):673, 2008.

[75] D. Yan, F. Vicini, J. Wong, and A. Martinez. Adaptive radiation therapy. *Physics in medicine and biology*, 42(1):123, 1999.

[76] L. Zollei, E. Grimson, A. Norbash, and W. Wells. 2D-3D rigid registration of X-ray fluoroscopy and CT images using mutual information and sparsely sampled histogram estimators. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 2, pages II–696. IEEE, 2001.

[77] Kelly H Zou, Simon K Warfield, Aditya Bharatha, Clare Tempany, Michael R Kaus, Steven J Haker, William M Wells III, Ferenc A Jolesz, and Ron Kikinis. Statistical validation of image segmentation quality based on a spatial overlap index: scientific reports. *Academic radiology*, 11(2):178–189, 2004.