UNIVERSITY OF CALIFORNIA

Los Angeles

Towards On-line Adaptive Therapy through the
Automation and Acceleration of Processes
on Graphics Processing Units

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Biomedical Physics

by

John Paul Neylon

2016

ABSTRACT OF THE DISSERTATION

Towards On-line Adaptive Therapy through the Automation and Acceleration of Processes on Graphics Processing Units

by

John Paul Neylon

Doctor of Philosophy in Biomedical Physics

University of California, Los Angeles, 2016

Professor Daniel Abraham Low, Chair

Adaptive therapies (ART) have potential for improving treatment efficacy, reducing unnecessary exposure of normal tissues, and improving patient quality of life. Ideally, every patient could receive on-line ART, fully optimizing the treatment to their daily anatomy as they lie on the treatment table. Additionally, daily on-line ART would allow reductions in the planned error margins by more certainly locating the tumor targets, providing another avenue for reducing exposure to normal tissues. To date, the computational complexity, labor, and time required to perform the additional tasks necessary for on-line ART has made it an

infeasible option for clinical implementation. Accelerating and automating these processes as much as possible will be imperative for clinical integration.

Towards this goal, software was developed for performing fast dose calculations, dose accumulation, contour propagation and analysis, deformable image registration (DIR) validation and error quantification, and biomechanical modeling. Each of these processes were accelerated for near real-time performance by parallelization and optimization for the architecture of graphics processing units (GPUs). Brief descriptions of the major contributions are given below.

A non-voxel-based dose convolution optimized for GPU architecture achieved over 4000x acceleration compared to a single-threaded implementation. Expanding this algorithm to a multi-GPU cloud-based implementation further increased the acceleration by a factor of two, despite the additional overhead associated with a distributed, cloud-based solution.

A DIR and dose accumulation framework was developed to track anatomical changes over the treatment course and estimate the actual delivered dose distribution. This framework was employed in retrospective studies to analyze the dose to the parotid glands for head-and-neck patients, and determine the feasibility of reducing error margins during planning.

A biomechanical modelling framework was developed to create patient-specific models from diagnostic imaging. Through GPU implementation, the high-resolution model maintained interactive framerates, for both linear elasticity and the subsequent evolution to hyper-elasticity. To validate the DIR algorithm employed in the dose accumulation framework, clinically realistic deformations were induced in patient-specific biomechanical models, which output simulated imaging volumes with known, ground-truth deformation vector fields.

Similar model-generated deformations supplied annotated training data for the development of a neural network able to infer a quantified error estimates for clinical DIR, requiring only

image similarity information as input. This methodology delivers a fully automated, fast technique to replace a process that was historically time-consuming, user-biased, and subject to small sample sizes.

The works presented focused on head-and-neck patients, but were developed with a general approach and the intent to expand to other sites. With future integration, these tools provide a foundation for building an automated, accelerated pipeline for clinical implementation of on-line ART.

The dissertation of John Paul Neylon is approved.

Ke Sheng

Patrick A Kupelian

Joseph M Teran

Anand Prasad Santhanam

Daniel Abraham Low, Chair

University of California, Los Angeles

2016

To my mother, Mary, and my father, Frank

TABLE OF CONTENTS

CHAPTER 1: Introduction to the Dissertation

CHAPTER 2: A Non-Voxel-Based Dose Convolution / Superposition Algorithm Optimized for Scalable GPU Architectures[1]

CHAPTER 3: Analytical Modeling and Implementation of a Multi-GPU Cloud-based Server (MGCS) Framework for Non-voxel-based Dose Calculation[2]

CHAPTER 4: Near Real-time Assessment of Anatomic and Dosimetric Variations for Head-and-neck Radiotherapy via a GPU-based Dose Deformation Framework[3]

CHAPTER 5: Feasibility of Margin Reduction for Level II and III Planning Target Volume in Head-and-neck Image-guided Radiotherapy - Dosimetric Assessment via a Deformable Image Registration Framework[4]

CHAPTER 9: Conclusion of the Dissertation

[1] A version of this chapter has been published as a manuscript in Medical Physics

[2] A version of this chapter has been submitted for review to the International Journal of Computer Assisted Radiology and Surgery

[3] A version of this chapter has been published as a manuscript in the International Journal of Radiation Oncology Biology Physics

[4] A version of this chapter has been published as a manuscript in Current Cancer Therapy Reviews

[5] This chapter is currently being prepared for submission to Medical Physics

LIST OF FIGURES

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 7

CHAPTER 8

CHAPTER 9

LIST OF TABLES

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 7

CHAPTER 8

ACKNOWLEDGMENTS

I would first like to thank my advisor, Dr. Anand Santhanam. I could not have asked for a more enthusiastic and supportive mentor. Anand has a unique talent for seeing potential in every project, and was always able to inspire, excite, and motivate me. This, I believe, is the most valuable tool that a researcher can be bestowed. Anand has become a friend and colleague in addition to a mentor, and I look forward to our continued work together.

However, the first professor I approached at UCLA was Dr. Ke Sheng. He was instantly engaging, and first sparked my interest in dose calculation algorithms. He also introduced me to Anand. So I would like to thank him for starting me down this path, and his continued support and guidance. I must also thank Ke for recommending me to Dr. Sharon Qi. Sharon has been one of my biggest proponents, and my work would not have been nearly as productive without her time and effort.

I would like to thank Dr. Daniel Low for his insight, Dr. Patrick Kupelian for his knowledge, and Dr. Joseph Teran for his expertise. At times, I was too focused on the computing aspects and minutiae of a project, and having the perspectives of three experts in their respective fields was invaluable.

I want to thank my parents, Mary and Frank, for their support and sacrifice. My mother first motivated me to pursue this field, and continues to be my inspiration. And my father is an unwavering source of strength and pride. I would also like to thank my sister, Delia, for keeping my ego in check.

Lastly, I would like to thank Floramae, for her love and support. Los Angeles has become home because of her more than anything, and I surely would have lost my mind long ago without her.

VITA

EDUCATION

B.S., Physics	Purdue University	2009
B.S., Mathematics	Purdue University	2009
M.S., Medical Physics	Purdue University	2010

AWARDS

Compute the Cure Fellowship	NVIDIA Foundation	2015

PEER REVIEWED PUBLICATIONS

1. K Hasse, **J Neylon**, K Sheng, and A Santhanam. "Systematic feasibility analysis of a quantitative elasticity estimation for breast anatomy using supine/prone patient postures," *Medical Physics* 43(3), p1299 (2016).

2. **J Neylon**, X. Sharon Qi, K Sheng, R Staton, J Pukala, R Manon, DA Low, P Kupelian, and A Santhanam. "A GPU-based high-resolution multi-level biomechanical head-and-neck model for validating deformable image registration," *Medical Physics* 42(1), p232 (2015).

3. X Sharon Qi, A Santhanam, **J Neylon**, Y Min, T Armstrong, K Sheng, R Staton, J Pukala, A Pham, DA Low, S Lee, M Steinberg, R Manon, A Chen, and P Kupelian. "Near real-time assessment of anatomic and dosimetric variations for head-and-neck radiotherapy via a GPU-based dose deformation framework," *International Journal of Radiation Oncology Biology Physics* 92(1), p415 (2015).

4. O Ilegbusi, B Seyfi, **J Neylon**, and A Santhanam. "Analytic inter-model consistent modeling of volumetric human lung dynamics," *Journal of Biomechanical Engineering* 137(10), p101005 (2015).

5. **J Neylon**, K Sheng, V Yu, D Low, P Kupelian, and A Santhanam. "A non-voxel-based dose convolution/superposition algorithm optimized for scalable GPU architectures," *Medical Physics* 41(10), p101711 (2014).

6. X Sharon Qi, **J Neylon**, S Can, R Staton, J Pukala, P Kupelian, and A Santhanam. "Feasibility of margin reduction for level II and III planning target volume in head-and-neck image-guided radiotherapy – dosimetric assessment via a deformable image registration framework," *Current Cancer Therapy Reviews* 10(4), p323 (2014).

7. Y Min, **J Neylon**, A Shah, S Meeks, P Lee, P Kupelian, and A Santhanam. "4D-CT lung registration using anatomy-based multi-level, multi-resolution optical flow analysis and thin-plate splines," *International Journal of Computer Assisted Radiology and Surgery* 9(5), p875 (2014).

8. B White, A Santhanam, D Thomas, Y Min, J Lamb, **J Neylon**, S Jani, S Gaudio, S Srinivasan, D Ennis, and D Low. "Modeling and incorporating cardiac-induced lung tissue motion in a breathing motion model," *Medical Physics* 41(4), p043501 (2014).

9. A Santhanam, T Dou, **J Neylon**, Y Min, P Kupelian, and K Sheng. "Multi-scale, multi-modal image integration for image-guided clinical interventions in the head-and-neck anatomy," *Studies in Health Technology and Informatics* 184, p380 (2013).

CONFERENCE PROCEEDINGS

1. A Santhanam, **J Neylon**, JD Eldredge, J Teran, E Dutson, and P Benharash. “GPU-based parallelized solver for large scale vascular blood flow modeling and simulations,” *Studies in Health Technology and Informatics* 220, p345 (2016).

2. **J Neylon**, X Sharon Qi, D Low, and A Santhanam. “Real-time hyper-elastic biomechanical models of head and neck anatomy for model guided multi-modal deformabe image registrations,” *18th International Conference on the Use of Computers in Radiation Therapy*. London, UK. June 2016;

3. X Sharon Qi, **J Neylon**, Y Yang, L Yang, A Santhanam, A Chen, and D Low. “MRI-guided soft tissue alignment for head-and-neck radiotherapy and margin adaption assessed by a GPU-based dose deformable registration framework,” *18th International Conference on the use of Computers in Radiation Therapy*. London, UK. June 2016.

4. **J Neylon**, K Hasse, K Sheng, and A Santhanam. “Modeling and simulation of tumor-influenced high resolution real-time physics-based breast models for model-guided robotic interventions,” *SPIE Proceedings 9786, Medical Imaging 2016: Image-Guided Procedures, Robotic Interventions, and Modeling, 97860X*. San Diego, CA. March 2016.

5. T Dou, Y Min, **J Neylon**, D Thomas, P Kupelian, and A Santhanam. “Fast simulated annealing and adaptive Monte Carlo sampling based parameter optimization for dense optical-flow deformable image registration of 4DCT lung anatomy,” *SPIE Proceedings 9786, Medical Imaging 2016: Image-Guided Procedures, Robotic Interventions, and Modeling, 97860N*. San Diego, CA. March 2016.

SELECT PRESENTATIONS

1. “Characterization of a Parameterized Image Similarity Cost Function for Automated Patient and Site Specific DIR Accuracy Optimizations,” *AAPM 58th Annual Meeting*. Washington, D.C. July 2016.

2. “Systematic assessment of a novel image similarity parameterization metric for fully automated quantification of deformable image registration accuracy,” *AAPM 57th Annual Meeting*. Anaheim, CA. July 2015.

3. “A GPU-Based Method for Validating Deformable Image Registration in Head and Neck Radiotherapy Using Biomechanical Modeling,” *AAPM 56th Annual Meeting*. Austin, TX. July 2014.

4. “A GPU framework for developing interactive high-resolution patient-specific biomechanical models,” *AAPM 56th Annual Meeting*. Austin, TX. July 2014.

5. “Simulating High-Resolution Biomechanical Head and Neck Model using a Multi-GPU Framework,” *NextMed / Medicine Meets Virtual Reality Conference*. Manhattan Beach, CA. February 2014.

6. “Establishing a quantitative relation between the spectrum-sampling rate on the dose distribution accuracy using a GPU based convolution/superposition algorithm,” *AAPM 55th Annual Meeting*. Indianapolis, IN. August 2013.

CHAPTER 1: Introduction to the Dissertation

MOTIVATION

Adaptive Radiotherapy

The National Cancer Institute estimated the number of new head-and-neck (HN) cancer cases in the U.S. for 2016 to be 48,330 [1]. These cancers are frequently aggressive in their biologic behavior, often presenting with multiple primary tumors. However, HN cancer is highly curable if detected early and treated precisely. Approximately 60% of HN cancer patients receive some form of radiation therapy (RT). A major consideration when delivering tumoricidal RT doses is the resultant normal tissue toxicity that results from radiation exposure of healthy anatomy surrounding the tumor target, which may lead to a reduction of bodily functions. Specifically for HN patients, normal tissue toxicity can cause dry mouth (xerostomia), inability to swallow (dysphagia), bone necrosis, tooth decay and more. In some cases, this exposure can lead to secondary cancers [2, 3].

Some degree of normal tissue toxicity is currently unavoidable from a planning perspective due to the error margins that are built into the treatment plan to account for the dynamic nature of the patient's anatomy over a course of treatment that may last several weeks. Factors such as weight loss, tumor regression, patient positioning and posture changes all contribute to changing patient anatomy from day to day [4].

Typical daily image guidance finds the rigid transformation between the planning kilovoltage (kV) computed tomography (CT) scan and the daily positioning imaging, which is often a different modality, such as megavoltage (MV) CT or cone-beam (CB) CT. Rigid transformations align bony anatomy or markers, ignoring internal physiological changes and soft tissue deformations. Non-rigid anatomy dynamics are accounted for by expanded error margins around the tumor targets incorporated during the treatment planning stage. Figure 1.1 shows a diagram of the established treatment planning volumes in radiotherapy. The gross tumor volume (GTV) defines the visible extents of the tumor, while the clinical target volume (CTV) expands around the GTV with margins of a few millimeters to encompass any microscopic disease not readily visible. Finally, the planning target volume (PTV) adds an additional margin to account for uncertainty in the daily anatomy.

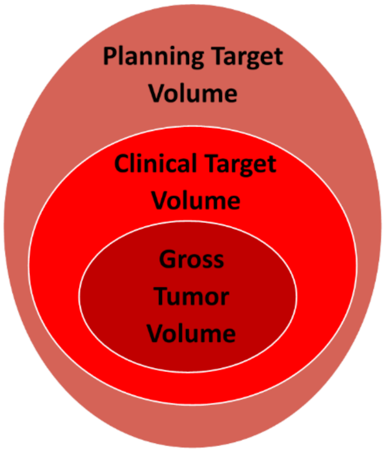


Figure 1.1. Illustration of radiotherapy treatment planning volumes as defined by the International Commission on Radiation Units & Measurements (Report 50 – Tx Volumes).

Theoretically, the rigid alignment should be sufficient to ensure proper tumor coverage due to the error margins included in the PTV. However, it has been reported that ignoring patient mis-alignments caused by non-rigid changes in patient posture and physiology can lead to under-dosing the tumor and over-irradiating the normal tissues [5, 6].

The risk of normal tissue toxicity and its resultant side effects can be reduced if the treatment plan is optimized using adaptive radiotherapy (ART) to account for the changes to patient anatomy over the treatment course. Several studies have shown that ART can provide significant dosimetric benefits for inter-fraction anatomic variations, as well as reduced toxicity, in the head-and-neck [7-10], as well as other cancer sites [11-14]. For instance, in 2009, Wu et al. compared strategies for adaptive re-planning and margin reduction for HN intensity modulated radiation therapy (IMRT), achieving dose sparing of the parotid glands (PG) to 30% [15]. They also showed that the results improved with the frequency and timeliness of the re-planning. In 2012, Capelle et al. performed a similar study using helical tomotherapy, where they observed volume changes up to 29% for the GTV, 17.5% for the PG, and overall weight loss of 3% [7]. With a single re-planning, the maximum dose delivered to the cord, and the mean dose to the PG was lowered over 1 Gray (Gy). More recently, in 2015, Castelli et al. assessed the impact of ART for sparing the PG, and how this lowered the risk of xerostomia [3]. They observed PG volume reductions of nearly 30%, and were able to reduce the mean PG dose by over 5 Gy with weekly adaptive re-planning.

Computational Challenges of On-line ART

While peers have investigated the potential benefits of ART, clinical implementations of ART are currently limited to off-line studies or require a significant amount of user intervention [7, 16]. Figure 1.2 shows a simplified flowchart for an off-line ART workflow. The ART workflow is inserted between fractions on a daily or weekly basis, analyzing the treatment to date and determining whether the patient would benefit from a new adaptive plan.

The computational challenges and increased manpower requirements of ART has inhibited full on-line capabilities, where the treatment is evaluated daily and plan adaptations computed, validated and enacted with the patient on the treatment table. In 2007, Xing et al. detailed the difficulty in moving from conventional RT to image guided RT (IGRT) to off-line ART and finally to on-line, image-guided ART [17]. They conjectured that the general processes stay largely the same along this evolution path for RT treatments, but the order, number of repetitions, and time scales between processes change dramatically. They went on to identify three major limiting factors recurring throughout the workflow. These were: (1) reliability: assessing and verifying the accuracy of each process; (2) integration: facilitating the communication of data between processes; (3) time: the effort and workforce required to complete all processes.

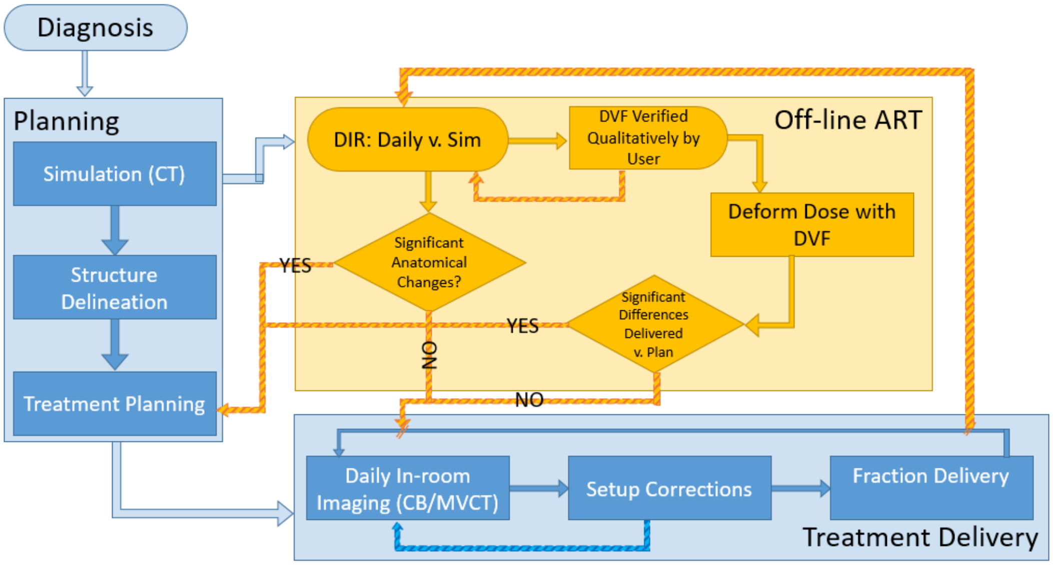


Figure 1.2. Simplified off-line ART workflow, where intervention is performed between fractions.

In order to feasibly work in a daily clinical workflow, the entire process could not take more than a few minutes to complete or clinical throughput would suffer. Therefore, the algorithms will need to be largely automated and highly accelerated to achieve the necessary performance. There are several possible bottlenecks where acceleration and automation could be applied to minimize the necessary time and labor. For instance, the iterative loop between DIR and verification would benefit from a fast, quantitative assessment of DIR performance. Initial DIR performance could be improved by an automated, patient-specific, validation study prior to the start of the treatment course. Fast dose calculations would facilitate dose accumulation, re-planning, and plan optimization. Dose accumulation also relies on DIR for contour propagation, which could benefit from site-specific DIR optimization. These are just a few processes where acceleration and automation greatly improves the feasibility of inclusion in the daily clinical workflow.

Xing et al. concluded their assessment of the current state of on-line ART and the computational challenges remaining to be addressed with the statement: “Automating radiotherapy processes through real-time adaptive image-guided strategies has the potential to make radiation treatments more accurate, efficient, and safe and should result in improved clinical outcomes.” The work in this dissertation was concentrated on the goals of automation and real-time performance, and harnessed the computational power of parallelization and the many-core architecture of graphics processing units to accomplish them.

BACKGROUND

General Purpose Computing on Graphics Processing Units (GPGPU)

Historically, central processing units (CPU) have progressed according to Moore’s Law, which states that the number of transistors on a chip can be doubled about every two years, effectively doubling the computational throughput. However, in recent years, transistors have neared their physical minimum at the nano-scale, leading CPU manufacturers to evolve from the single core to multiple core processors to continue improving compute performance. In the past decade or so this concept of multi-threaded processing has led to an explosion of general purpose computing on graphics processing units (GPUs).

GPU hardware was originally developed to accelerate computer graphics tasks, such as texture mapping and ray-tracing, which are considered “embarrassingly parallel.” These are tasks that have little or no inter-dependence or need for communication. Offering hundreds of processing cores on a single chip, GPUs are better suited for parallel processing of massive data sets that are commonly encountered for scientific research.

The latest GPU hardware architecture available from NVIDIA on the GeForce GTX Titan X, with compute capability 5.2, can launch nearly 50,000 threads in parallel, distributed across 24 multiprocessors, each of which contains 128 processing cores. Threads are launched in warps of 32, and execute the same compute kernel simultaneously. The GPU’s task scheduler distributes blocks of threads between the multiprocessors, attempting to utilize as many cores as possible and maximize GPU occupancy. Running at 7 gigabits per second (Gbps), the Titan X card can theoretically deliver 6600x10^9 floating point operations per second (6600 GFlops) of computational power for single precision arithmetic, and 206 GFlops for double precision. This is nearly an order of magnitude more computing power compared to the latest generation of CPUs from Intel, which have been reportedly clocked around 500 GFlops for single precision operations [18].

As the computational capabilities of GPUs continue to improve, the performance bottlenecks have shifted from hardware considerations such as data transfer bandwidth, to software and code design. Memory capacity has become less of a concern in recent years, with high end commercial chips, such as the Titan X, offering up to 12 gigabytes (GB) of random access memory (RAM). However, there are substantial design considerations when porting an algorithm from a CPU to GPU architecture.

GPU architecture has a unique memory hierarchy with variable data retrieval speeds and scopes [19, 20]. There are several tiers in the memory hierarchy of the GPU, each with different access latencies, scopes, and capacities. Figure 1.3 shows a schematic of a typical GPU and the scope of its different memory spaces.

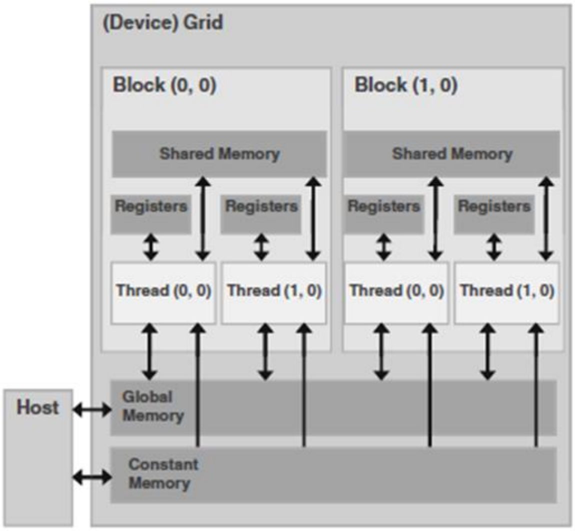


Figure 1.3. GPU memory hierarchy. This figure was originally published in D. Kirk and W. Hwu's *Programming Massively Parallel Processors*, in 2010 [19].

On-chip memories include shared memory and thread registers. These memory spaces provide high speed access of frequently called variables, but are limited in scope and capacity.

Registers are small caches private to individual threads, and survive only for the length of the kernel. This information cannot be shared between threads or thread blocks unless it is written to a higher tier of memory. Shared memory is similar, though it is larger and can be accessed by all threads in a single block. Blocks cannot access the shared memory space of other blocks, but the threads within a block can communicate variables through this space. Shared memory also survives only for the length of the kernel launch, so to preserve the information on shared memory, it must be transferred to the global space. The largest memory space, dubbed the 'global memory', is accessible by every block of threads, but accessing the data can have latencies of several hundred clock cycles. Shared memory offers access speeds 100-150 times faster than global memory, but its scope is limited and the size is extremely small compared to global memory. The Titan X has a limit of 49,152 bytes of shared memory per block. Additionally, the host (CPU) only has read and write access to the global and constant memory of the device (GPU), and the device only has read access to constant memory locations. This allows short latency and high bandwidth data transfer where all threads can access the same constant memory address.

The pattern in which the memory is accessed on the GPU also forms an important design consideration. As was mentioned above, accessing data from global memory incurs a fetch latency of several hundred clock cycles. However, if all the threads of a block make a call to global memory synchronously and access a contiguous block of global memory, this is considered a coalesced memory access and costs only one fetch, instead of incurring a fetch for each thread making a call. This alone can easily accelerate a compute kernel by two times or more. Sometimes, data access patterns are not predictable. In this case, the data in global memory can be assigned as texture memory, which makes it read-only to the GPU threads. This provides several advantages, including two-dimensional caching and hardware intrinsic bilinear interpolation.

In order to utilize the GPU to maximize performance capacity, memory latencies must be minimized and throughput maximized such that each thread always has a task to perform. But GPU memory spaces are just one design consideration when developing for GPU architecture. The parallelization strategy of the algorithm must also be determined by analyzing data inter-dependencies and sequencing. Another is the development environment.

NVIDIA introduced CUDA in 2006. CUDA is a general-purpose parallel computing platform and programming model, which offers low and high level application programming interfaces (APIs) that integrate with C, C++, and Fortran programming languages. Accelerated C/C++ libraries and extensions are included in the CUDA software development kit (SDK), along with a low level virtual machine compiler and compiler directives. The work in this dissertation was developed in C/C++, utilizing the CUDA SDK extensively, as well as the accelerated libraries, in a Linux environment running the Ubuntu operating system.

SPECIFIC AIMS

The ultimate goal of this dissertation was to facilitate on-line adaptive radiotherapy in the daily clinical workflow through the development of automated and accelerated processes. This stems from the hypothesis that an automated framework for accurately tracking anatomy, computing the accumulated dose delivered and reporting dosimetric endpoints for critical structures in near real-time will be vital for enabling on-line ART. As the entire adaptive re-planning process would need to finish while the patient waits on the treatment table, I focused on developing GPU-accelerated tools with near real-time performance that require as little user intervention as possible. The following specific aims focus on developing key enabling technologies for on-line ART:

Specific Aim 1 (SA1): Develop a fast dose convolution/superposition algorithm optimized for GPU architecture, with scalability for distributed workloads.

Specific Aim 2 (SA2): Develop a framework for fast DIR and dose accumulation estimation, with an automated methodology for quantifying the DIR performance.

Specific Aim 3 (SA3): Develop a framework for generating patient-specific, biomechanical models with the ability to reproduce clinically realistic deformations for the purpose of generating ground truth data for clinical DIR validation.

OVERVIEW

Chapters 2 through 8 consist of edited manuscripts produced from the core projects of this dissertation that have been published, are under review, or are currently being prepared for submission. As such, the background material for each project was not included in this chapter of the dissertation, since each chapter contains a thorough introductory section discussing the impetus for the project, prior work in the field, and the current state of the art.

SA1 is addressed in chapters 2 and 3. Chapter 2 details the development and optimization of a non-voxel-based (NVB) dose convolution/superposition algorithm, and discusses in depth the major design considerations for GPU programming. Chapter 3 extends the NVB dose algorithm to a multi-GPU cloud-based server (MGCS) framework, detailing methods and optimization techniques for scaling across multiple GPUs on a single machine, and multiple machines. Chapters 4, 5, and 8 address SA2. Chapter 4 presents a retrospective study performed using the framework for fast DIR and dose accumulation. Chapter 5 again utilizes the DIR and dose accumulation framework to revisit the data produced in chapter 4, to test the feasibility of reducing the error margins by comparing dosimetric endpoints. Chapter 8 describes the development of a methodology for automated quantification of DIR performance by parameterizing image similarity metrics.

SA3 is addressed by chapters 6 and 7. Chapter 6 describes the development of a framework for instantiating patient-specific, interactive biomechanical models from planning kVCTs, and a methodology for producing clinically realistic ground truth deformations to validate the DIR algorithm employed in the dose accumulation framework of chapters 4 and 5. Chapter 7 presents a further sophistication of the biomechanical modelling framework from chapter 6, including the incorporation of a hyper-elastic material model to more accurately characterize soft tissue response for large deformations.

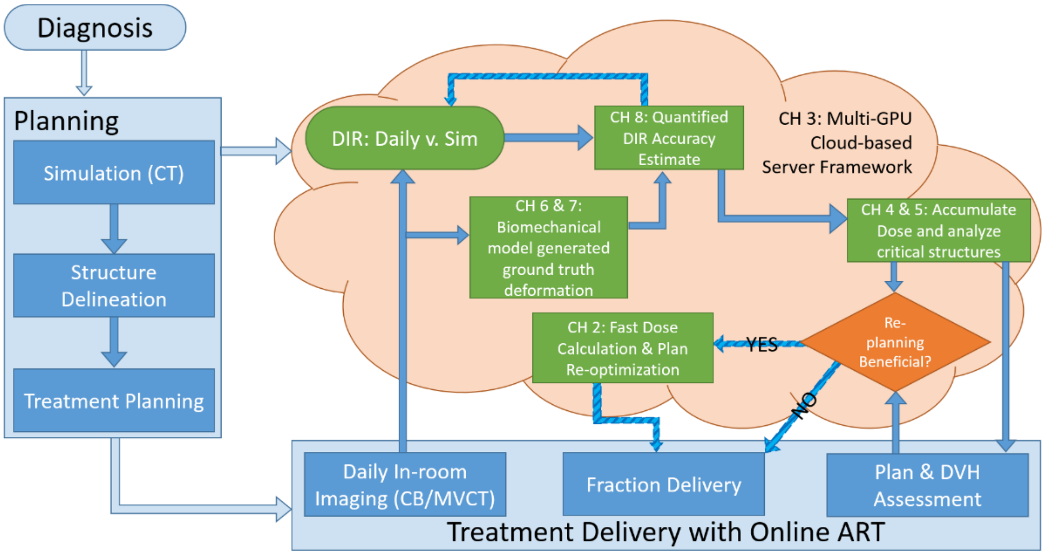

Diagnosis
Planning
Simulation (CT)
Structure Delineation
Treatment Planning
DIR: Daily v. Sim
CH 8: Quantified DIR Accuracy Estimate
CH 3: Multi-GPU Cloud-based Server Framework
CH 6 & 7: Biomechanical model generated ground truth deformation
CH 4 & 5: Accumulate Dose and analyze critical structures
CH 2: Fast Dose Calculation & Plan Re-optimization
YES
Re-planning Beneficial?
NO
Daily In-room Imaging (CB/MVCT)
Fraction Delivery
Plan & DVH Assessment
Treatment Delivery with Online ART

Figure 1.4. A potential workflow for on-line ART, combining the works of this dissertation. Here the registration enters an automated optimization loop, where performance is assessed quantitatively based on the works in chapters 6, 7, and 8. The deformation vector field is then sent to the dose accumulation and plan assessment framework detailed in chapters 4 and 5, producing dose volume histograms and comparing the estimated delivered dose with the plan. After treatment assessment, the plan is adapted using the fast dose calculation described in chapter 2. Lastly, all these tools are instantiated on a multi-GPU cloud-based framework, as described in chapter 3.

Figure 1.4 modifies the ART flowchart from figure 1.2, and illustrates how these wide-ranging projects may be combined in the future within a potential on-line ART flowchart. The computationally heavy tasks of on-line ART are performed remotely, scaled across a multi-GPU server framework using the strategy detailed in chapter 3, and running concurrently with therapists' normal tasks. The DIR is automatically optimized using methodologies of chapters 6, 7, and 8. The DIR would then be used by the dose accumulation framework described in chapters 4 and 5 to assess the treatment to date, producing contour-specific endpoints and DVH data, and determine whether plan adaptation is necessary. Lastly, the fast dose calculation engine described in chapters 2 and 3 could be used to re-calculate the dose on the daily anatomy, re-planning, and plan optimization.

This flowchart is revisited in chapter 9, where the conclusions of this dissertation are presented, and future avenues of pursuit for each of the projects are discussed.

REFERENCES

[1] *Cancer facts & figures 2016*. 2016, American Cancer Society: Atlanta: American Cancer Society.

[2] Bentzen, S. M., Constine, L. S., Deasy, J. O., Eisbruch, A., Jackson, A., Marks, L. B., Ten Haken, R. K., and Yorke, E. D., "Quantitative analyses of normal tissue effects in the clinic (quantec): An introduction to the scientific issues.," International Journal of Radiation Oncology Biology Physics, S3-S9 (2010).

[3] Castelli, J., Simon, A., Louvel, G., Henry, O., Chajon, E., Nassef, M., Haigron, P., Cazoulat, G., Ospina, J. D., Jegoux, F., Benezery, K., and de Crevoisier, R., "Impact of head and neck cancer adaptive radiotherapy to spare the parotid glands and decrease the risk of xerostomia," Radiation Oncology 10(1), 6 (2015).

[4] Castadot, P., Lee, J. A., Geets, X., and Gregoire, V., "Adaptive radiotherapy of head and neck cancer," Seminars in Radiation Oncology 20(2), 84-93 (2010).

[5] Liu, Erickson, Peng, and Li, "Characterization and management of interfractional anatomic changes for pancreatic cancer radiotherapy," International Journal of Radiation Oncology Biology Physics 83(3), e423-e429 (2012).

[6] Bujold, A., Craig, T., Jaffray, D., and Dawson, L., "Image-guided radiotherapy: Has it influenced patient outcomes?," Seminars in Radiation Oncology 22(1), 50-61 (2012).

[7] Capelle, L., Mackenzie, M., Field, C., Parliament, M., Ghosh, S., and Scrimger, R., "Adaptive radiotherapy using helical tomotherapy for head-and-neck cancer in definitive and postoperative settings: Initial results," Clinical Oncology 24(3), 208-215 (2012).

[8] Foroudi, F., Wong, J., Kron, T., Rolfo, A., Haworth, A., Roxby, P., Thomas, J., Herschtal, A., Pham, D., Williams, S., Tai, K. H., and Duchesne, G., "Online adaptive radiotherapy for muscle-invasive bladder cancer: Results of a pilot study," International Journal of Radiation Oncology Biology Physics 81(3), 765-771 (2011).

[9] Zeidan, O. and Huddleston, A. J., "A comparison of soft-tissue implanted markers and bony anatomy alignments for image-guided treatments of head and neck cancers.," International Journal of Radiation Oncology Biology Physics (2009).

[10] Zeidan, O. and Langen, K. M., "Evaluation of image-guidance protocols in the treatment of head and neck cancers," International Journal of Radiation Oncology Biology Physics 67(3), 670-677 (2007).

[11] Lindegaard, J., Fokdal, L., Nielsen, S., Juul-Christensen, J., and Tanderup, K., "Mri-guided adaptive radiotherapy in locally advanced cervical cancer from a nordic perspective," Acta Oncologica 52(7), 1510-1519 (2013).

[12] Nijkamp, J., Marijnen, C., Herk, M. V., Triest, B. V., and Sonke, J., "Adaptive radiotherapy for long course neo-adjuvant treatment of rectal cancer," Radiotherapy and Oncology 103(3), 353-359 (2012).

[13] Schwartz, D., Garden, A., Thomas, J., Chen, Y., Zhang, Y., Lewin, J., Chambers, M., and Dong, L., "Adaptive radiotherapy for head-and-neck cancer: Initial clinical outcomes from a prospective trial," International Journal of Radiation Oncology Biology Physics 83(3), 986-993 (2012).

[14] Tuomikoski, L., Collan, J., Keyrilainen, J., Visapaa, H., Saarilahti, K., and Tenhunen, M., "Adaptive radiotherapy in muscle invasive urinary bladder cancer - an effective method to reduce the irradiated bowel volume," Radiotherapy and Oncology 99(1), 61-66 (2011).

[15] Wu, Q., Chi, Y., Chen, P. Y., Krauss, D. J., Yan, D., and Martinez, A., "Adaptive replanning strategies accounting for shrinkage in head and neck imrt," International Journal of Radiation Oncology Biology Physics 75(3), 924-32 (2009).

[16] Veiga, C., McClelland, J., Moinuddin, S., Lourenco, A., Ricketts, K., Annkah, J., Modat, M., Ourselin, S., D'Souza, D., and Royle, G., "Toward adaptive radiotherapy for head and neck patients: Feasibility study on using ct-to-cbct deformable registration for "dose of the day" calculations," Medical Physics 41(3), 031703 (2014).

[17] Xing, L., Siebers, J., and Keall, P., "Computational challenges for image-guided radiation therapy: Framework and current research," Seminars in Radiation Oncology 17(4), 245-257 (2007).

[18] NVIDIA, *Cuda c programming guide*. 2015: www.nvidia.com.

[19] Kirk, D. and Hwu, W.-m., [Programming massively parallel processors], Burlington, MA: Elsevier, Inc., (2010).

[20] Sanders, J. and Kandrot, E., [Cuda by example], Boston, MA: Pearson Education, Inc., (2011).

CHAPTER 2: A Non-Voxel-Based Dose Convolution / Superposition Algorithm Optimized for Scalable GPU Architectures

A version of this chapter has been published as a manuscript: Med Phys, Vol. 41, No.10, 101711, 2014. doi: 10.1118/1.4895822

ABSTRACT

Purpose. Real-time adaptive planning and treatment has been infeasible due in part to its high computational complexity. There have been many recent efforts to utilize graphics processing units (GPUs) to accelerate the computational performance and dose accuracy in radiation therapy. Data structure and memory access patterns are the key GPU factors that determine the computational performance and accuracy. In this paper, we present a non-voxel-based approach to maximize computational and memory access efficiency and throughput on the GPU.
Methods. The proposed algorithm employs a ray-tracing mechanism to re-structure the 3D data sets computed from the CT anatomy into a non-voxel-based framework. In a process that takes only a few milliseconds of computing time, the algorithm re-structured the datasets by ray-tracing through pre-calculated CT volumes to re-align the coordinate system along the convolution direction, as defined by a zenithal and azimuthal angle. During the ray-tracing step, the data were resampled according to radial sampling and parallel ray spacing parameters, making the algorithm independent of the original CT resolution.

The non-voxel-based algorithm presented in this paper also demonstrated a trade-off in computational performance and dose accuracy for different coordinate system configurations. In order to find the best balance between the computed speed up and the accuracy, we employed an exhaustive parameter search on all the sampling parameters that defined the coordinate system configuration: the zenithal, azimuthal, and radial sampling of the convolution algorithm, as well as the parallel ray spacing during ray-tracing. The angular sampling parameters were varied between 4 and 48 discrete angles, while both radial sampling and parallel ray spacing were varied from 0.5 to 10 mm. The gamma distribution analysis method (□) was used to compare the dose distributions using 2% and 2mm dose-difference and distance-to-agreement criteria, respectively. Accuracy was investigated using three distinct phantoms with varied geometries and heterogeneities and on a series of 14 segmented lung CT datasets. Performance gains were calculated using three 256 mm cube homogenous water phantoms, with isotropic voxel dimensions of 1 mm, 2 mm, and 4 mm.

Results. The non-voxel-based GPU algorithm was independent of the data size, and provided significant computational gains over the CPU algorithm for large CT data sizes. The parameter search analysis also showed that the ray combination of 8 zenithal and 8 azimuthal angles, along with 1 mm radial sampling and 2 mm parallel ray spacing maintained dose accuracy with greater than 99% of voxels passing the □ test. Combining the acceleration obtained from GPU parallelization with the sampling optimization, we achieved a total performance improvement factor of >175,000 when compared to our voxel-based ground truth CPU benchmark, and a factor of 20 compared with a voxel-based GPU dose convolution method.

Conclusions. The non-voxel-based convolution method yielded substantial performance improvements over a generic GPU implementation, while maintaining accuracy as compared to a CPU computed ground truth dose distribution. Such an algorithm can be a key contribution towards developing tools for adaptive radiation therapy systems.

INTRODUCTION

Radiotherapy has seen a major push towards treatment plans that are tailored to the patient and adapted to their radiation response[1-4]. Ignoring inter- and intra-treatment changes in tumor size and position can lead to target under-dosing and excessive exposure of healthy tissue[3, 5]. Real-time adaptive therapy has been infeasible due in part to the time and computational effort required for such tasks[6].

In recent years, graphics processing units (GPU) have gained widespread use in scientific computing, due to its massive parallelization, allowing thousands of times more floating point operations per second than a typical CPU[7, 8]. There are several hurdles along the path of a GPU implementation, but their acceleration capabilities have made radiation oncology challenges such as live tumor tracking and real-time dose estimations into realistic possibilities[9].

Advantages of using GPUs for dose calculations have been previously examined, specifically in regard to the convolution/superposition algorithm. Three independent groups have implemented the superposition/convolution onto GPU architecture. Hissoiny et al. reported acceleration of 10-20x in 2009, and later improved to nearly 30x when compared to an optimized commercial CPU implementation[10, 11].

In 2011, GPU acceleration was pushed above 100x compared to an optimized dual core CPU[12, 13]. Dose calculation accuracy of GPU and CPU implementations were compared by using 48 zenithal angles and 96 azimuthal angles. The accuracy, calculated as the percent dose difference between corresponding voxels relative to the maximum dose, agreed to within 2-5%. While these methods employed voxel-based calculations, Chen et al.[14, 15] employed a non-voxel based broad beam framework first proposed by Lu[16] to perform the calculations prior to convolution, but did not extend it to the actual convolution. Acceleration factors of 1000-

3000 were reported using their exponential kernel on GPU compared to a tabulated kernel on CPU.

As the computational capabilities of GPUs continue to improve, the performance bottlenecks have shifted from hardware considerations such as data transfer and maximum number of parallel threads, to software and code design considerations. The GPU architecture has a unique memory hierarchy with variable data retrieval speeds and scopes[7, 8]. In addition, the pattern in which the memory is accessed on the GPU also forms an important design consideration. To fully utilize the potential computing power of the GPU, the memory design aspects must be considered, requiring approaching old problems from new viewpoints. Convolution/superposition still provides the best compromise between speed and accuracy when performing dose calculations in heterogeneous materials. However, because of its inherent memory access pattern the convolution process is performance-limiting when trying to port the algorithm to the GPU, specifically the spherical sampling pattern about the point of interest.

In this paper, a GPU accelerated superposition/convolution is presented that employs an improved memory assignment optimization. Specifically, a non-voxel-based (NVB) GPU-accelerated superposition/convolution algorithm and its dependence on sampling parameters are presented. Converting the dose convolution calculation to new coordinate systems aligned along each convolution direction allows for fully optimized memory access patterns along each step of the algorithm and provides a significant computational speed-up. We also introduce a fourth sampling parameter, the spacing between parallel rays when resampling for the NVB coordinate system, alongside the traditional spherical sampling variables of the convolution algorithm. Utilization of greater sampling rates prolongs computational times and so was previously avoided for CPU based dose calculation frameworks. Characterizing the accuracy and performance effects of varying coordinate system parameters allows greater control over the convolution, further optimizing the algorithm.

MATERIALS AND METHODS

In this section, we first describe the collapsed cone convolution (A). It is followed by a discussion on the convolution sampling, how it dictates the memory access patterns, and the resultant performance considerations (B). We then present the non-voxel-based algorithm (C), and detail the experiments to quantify the non-voxel-based algorithm's accuracy (D), and compare its performance to a voxel-based CPU algorithm and a generic GPU implementation (E).

Collapsed Cone Convolution/Superposition Algorithm

Collapsed cone convolution/superposition (CCCS) has been well documented[17-19], in this section we present a brief review of its mathematics.

TERMA calculation. In order to calculate the total energy released in matter, or TERMA, the equivalent depth in water must be known for each voxel in the target. For calculation purposes, the beam was assumed to be originating from a point source 1 meter above the isocenter. Siddon's ray-tracing algorithm was ported to GPU architecture for this task[19]. In order to compute the primary energy deposition, the attenuation path of each ray was corrected for density heterogeneities. This effective radiological path length in water was calculated from source to voxel, by summing the contributions of each voxel along the ray path,

$$d_i = \sum_j l_j \rho_j, \qquad (1)$$

where *i* was the point of interaction, *j* was the voxel the ray intersected, l_j was the intersection length of the ray and the voxel, ρ_j was the voxel density relative to water and therefore unitless.

Equations 2 and 3 show the discrete formulas for the TERMA with a beam hardening correction, summed over the discretized energy spectrum, *E*,

$$T(i) \ = \ \sum_E \Psi_E \mu_E e^{-\mu_E d_i}, \tag{2}$$

$$T'(i) = \left(\frac{\sum_E \Psi_E \mu_{en,E} e^{-\mu_E d_i} / T(i)}{\sum_E \Psi_E \mu_{en,E} e^{-\mu_E d_0} / T(0)}\right) * T(i), \tag{3}$$

where *i* was the point of interaction, Ψ was the energy fluence, μ was the mass attenuation coefficient, and μ_{en} was the mass energy absorption coefficient. Equation 3 shows the correction factor for beam hardening using the unattenuated values[20]. The attenuation coefficients were drawn from the National Institute of Standards and Technology database[21].

Cumulative-cumulative kernel generation. The CCCS dose distribution was calculated by convolving a poly-energetic cumulative-cumulative dose deposition kernel (CCK) with the TERMA volume computed using equation (2) and (3)[17, 18]. The kernel files were pre-computed, Monte Carlo generated, mono-energetic differential deposition distribution kernel (*DK*) about a point interaction. For each geometric location, the kernel files described the energy dispersal due to the type of interaction (T): primary interaction, first scatter, second scatter, multiple scatter, and bremsstrahlung/annihilation[22]. To create mono-energetic cumulative kernels (*CK)*, the initial differential kernels were summed over the interaction type, and integrated over the spherical sampling space[23]. The cumulative kernels were then integrated over the sampling space again, and then summed over the energy spectrum (E) according to their spectrum weight (*w_E*), constructing a single poly-energetic cumulative-cumulative kernel, (CCK)[23, 24]

$$CK(\theta, \varphi, r) = \ \int (\sum_T DK_T(\theta, \varphi, r)) \ dr \tag{4}$$

$$CCK(\theta, \varphi, r) = \ \int w_E \ (\int CK(\theta, \varphi, r) \ dr) dE. \tag{5}$$

CCK dose convolution. The superposition method was employed to scale the kernel,

$$Dose(v) = \int T'(v')CCK(\bar{\rho}_{v-v'} * v - v')dv'; \; where \; \int dv' = \iiint d\varphi d\theta dr, \qquad (6)$$

where *v* was the interaction point, *v'* was the voxel being sampled, $\bar{\rho}_{v-v'}$ was the heterogeneity correction applied to the kernel, r was the radial component, θ was the zenith angle, and φ was the azimuthal angle. The CPU algorithm tackled this process using nested loops, which cycled through each voxel, *v*, within the beam and then sampled the surrounding volume ($\iiint v'dV$), before moving on to the next voxel.

Convolution Sampling and Memory Access Patterns

The discretized convolution algorithm employed during the CCCS calculations (equation 6) required spherical sampling about the voxel of interest and summing the dose contributions of the surrounding volume. In practice, the dose at the point of interaction was calculated by summing the contributions of the discretely-sampled surrounding volume according to these three parameters: the number of zenithal angles (Θ), the number of azimuthal angles (Φ), and the size of the radial increment (P). The number of sampling points and the computation time were linearly related to Θ and Φ, and inversely related to P. The limit of sampling resolution was set by the kernel file parameters. The dose deposition kernels were segmented into 24 concentric circles with varying radii from 0.1 to 60 cm, and 48 zenithal segmentations equally spaced from 0 to 180 degrees. This effectively created a ceiling to the zenithal and radial sampling during convolution. Azimuthal sampling was limitless in theory, because the CCK was computed for a homogenous material. This resulted in symmetric dose deposition about the azimuth, and therefore the information was only recorded for two dimensions. However, when applying the heterogeneity correction, azimuthal sampling could have a profound effect on computation accuracy.

Generic GPU implementation. A generic method to parallelize the algorithm was developed initially, similar to the first published GPU implementation of the convolution/superposition algorithm[10, 12]. This simplistic approach launched a GPU function for each zenithal and azimuthal angle combination, unrolled the outermost loops which cycle through each voxel, and convolved them simultaneously. Each voxel within the volume was assigned a thread and traced a ray from that voxel in the direction specified by the zenithal angle, *Θ*, and the azimuthal angle, *Φ*, sampling the density and TERMA at radial intervals of *P* and applying the CCK, scaled by the density for heterogeneity. For the ground truth sampling parameters of 48/48, this amounted to 2304 function launches.

Performance considerations and bottlenecks. The conventional GPU algorithm presented several hurdles when attempting to optimize memory access patterns for GPU architecture. The GPU contained several memory types with varying scopes and access speeds. Global memory had the largest capacity but also had the greatest latency when accessing data, typically between 400-600 clock cycles. Shared memory offered access speeds 100-150 times faster than global memory, but had scope limited to a single block of threads, and a much reduced capacity[8]. Global access speeds could approach shared access speeds if the memory fetches were coalesced. This facilitated a group of adjacent threads to simultaneously read from a group of adjacent memory addresses in the global memory space. The compiler will then combine these into a single larger memory fetch, greatly reducing the latency[7]. Another design limitation is that the GPU's shared memory cannot be written into directly by the CPU. The threads of the block must read in the data from global memory first, and fill the shared memory space. Therefore, the most efficient way to attack the convolution is to organize the threads along the convolution ray direction, utilize coalesced global memory fetches to write into shared memory, and then use shared memory to perform the convolution.

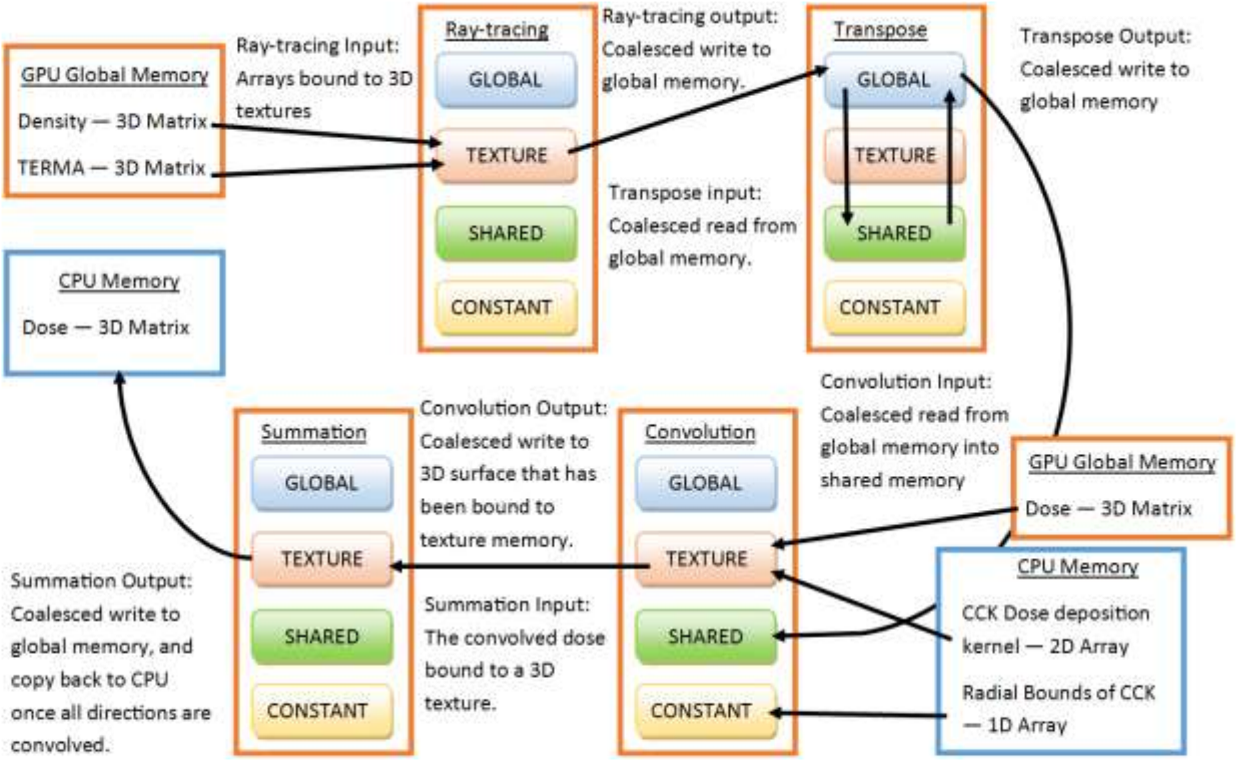

GPU Global Memory
Density — 3D Matrix
TERMA — 3D Matrix
Ray-tracing Input: Arrays bound to 3D textures
Ray-tracing
GLOBAL
TEXTURE
SHARED
CONSTANT
Ray-tracing output: Coalesced write to global memory.
Transpose input: Coalesced read from global memory.
Transpose
GLOBAL
TEXTURE
SHARED
CONSTANT
Transpose Output: Coalesced write to global memory
CPU Memory
Dose — 3D Matrix
Summation
GLOBAL
TEXTURE
SHARED
CONSTANT
Convolution Output: Coalesced write to 3D surface that has been bound to texture memory.
Convolution
GLOBAL
TEXTURE
SHARED
CONSTANT
Convolution Input: Coalesced read from global memory into shared memory
GPU Global Memory
Dose — 3D Matrix
CPU Memory
CCK Dose deposition kernel — 2D Array
Radial Bounds of CCK — 1D Array
Summation Output: Coalesced write to global memory, and copy back to CPU once all directions are convolved.
Summation Input: The convolved dose bound to a 3D texture.

Figure 2.1. GPU memory flowchart for the NVB dose convolution algorithm.

The problem here is that coalesced access is only possible in one direction, while the convolution rays can have any arbitrary direction as defined by Θ and Φ. Texture memory is located in the global memory space, but is cached for locality and also provides an intrinsic linear interpolation in three dimensions. This makes it ideal when coalesced accesses are not possible but the memory reads are patterned predictably. However, it is read-only unless it is created using a specialized array that can be bound to a surface object. This feature is only available on more recent generations of devices with compute capabilities of 3.0 or higher.

Non-Voxel-Based Algorithm

In this section we present the framework of our non-voxel-based algorithm. To take advantage of the different memory spaces and maximize efficiency, we split the convolution into four components: ray-tracing, transposition, line convolution, and summation. These four steps were performed for every zenithal direction less than or equal to 90 degrees, and every

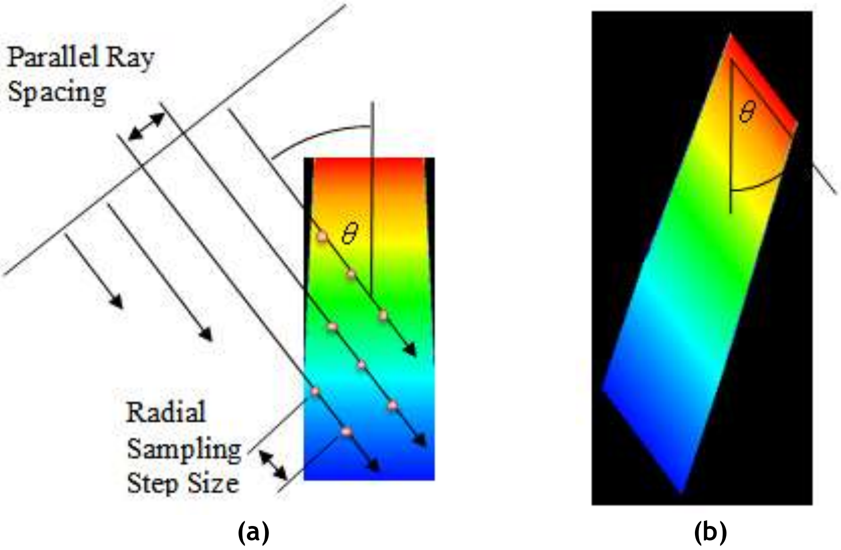
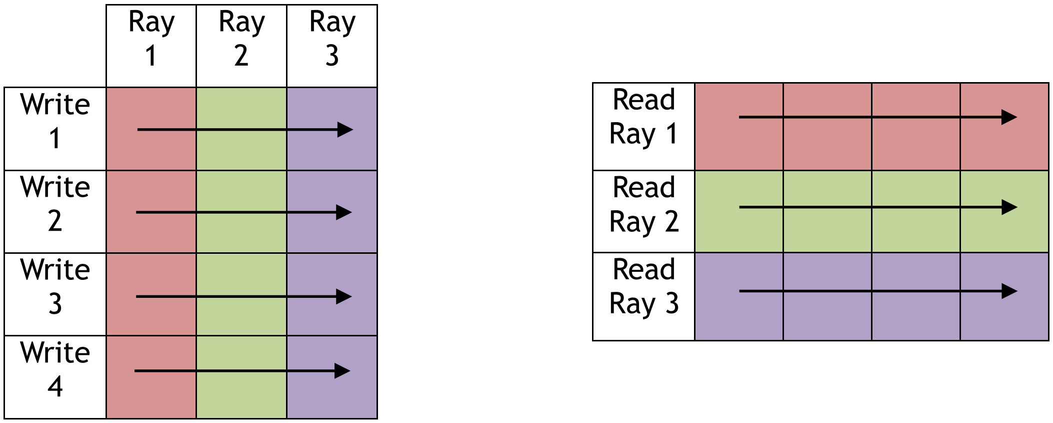


Figure 2.2. NVB Algorithm Ray-tracing. (a) An example TERMA map with the convolution angle determined by θ. (b) Ray-tracing. The TERMA resampled along the convolution direction.

azimuthal direction. Figure 2.1 illustrates the movement of data between GPU memory spaces during the process.

Ray-tracing. We first converted the density and TERMA data volumes from voxelized Cartesian coordinates into a non-voxel based coordinate system aligned with the convolution ray direction. To do this, the density and TERMA data (already residing in the GPU's global memory from the TERMA calculation) were bound to 3D textures in the GPU's texture memory. It was then possible to trace through the volumes with a grid of parallel rays, equally spaced by a distance, Δ. During ray tracing, the volumes were sampled at intervals equal to the pre-defined radial step size of the convolution, P. Figure 2.2(a) illustrates the process. The parallel rays were incident on the TERMA map at a zenithal angle, θ. The parallel rays were evenly spaced in a 2 dimensional grid.

The spacing of the parallel rays is the fourth sampling parameter introduced by the NVB convolution method. As the rays traced through the volume, they sampled the TERMA data at regular intervals defined by the radial sampling step size.



(a) Ray-tracing Write Pattern (b) Line Convolution Read Pattern

Figure 2.3. NVB Algorithm Transposition and Coalesced Memory Access. The different colors represent different rays as they trace through the volume. The nature of the ray-tracing algorithm only allows a coalesced memory write by assigning adjacent memory locations to adjacent rays, as in (a). However, in order to perform a coalesced read into shared memory for the line convolution, adjacent memory locations must represent adjacent sampling points along the same ray, as illustrated in (b).

By utilizing texture memory and its intrinsic linear interpolation to resample the TERMA and density, the computational complexity remained at a maximum of O(p), where p was the total number of sampling points as dictated by the radial step size and parallel ray spacing parameters and cropped to the size of the field plus penumbra. The number of rays was dynamically allocated depending on the data size and the angle of incidence. The new non-voxel based data volumes were written back into global memory using a coalesced write, such that adjacent memory addresses represent adjacent rays. Figure 2.2(b) displays the TERMA map from figure 2.2(a) in the new non-voxel-based coordinate system.

Transposition. These data volumes were transposed to facilitate a coalesced memory read where adjacent memory addresses represented the sampling points along a single ray. The memory access patterns of the ray-tracing write and the line convolution read are illustrated in figure 2.3. This was done by doing coalesced reads into shared memory tiles, transposing the tiles, and performing a coalesced write back into global memory. The spatial complexity of the transposition was also O(p).

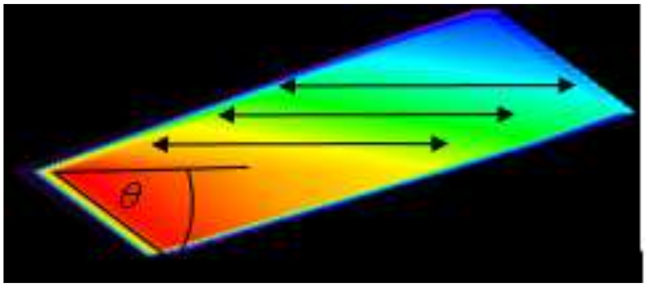


Figure 2.4. NVB Algorithm Line Convolution. The figure shows the transpose of the reformatted TERMA map in figure 1(b). The line convolution was performed along each ray as indicated by the arrows.

Line convolution. Now that the data were aligned along the convolution direction, a simple line convolution was performed [25]. The data were again loaded into shared memory using a coalesced read. Each thread in the block represented one sampling point along a given ray. Figure 2.4 illustrates the line convolution, displaying the result of convolving the transposed TERMA map.

Each GPU thread then stepped away from itself in both directions along the ray, accumulating the dose by multiplying the TERMA from each voxel with the CCK, and applying the heterogeneity corrections by sampling the density. The value of the CCK was calculated by first comparing the effective radiological distance of the current voxel to an array of the radial boundaries of the CCK in the GPU's constant memory. The CCK was loaded as a 2D array into the GPU's texture memory to take advantage of the intrinsic linear interpolation on the GPU, which kept the computational complexity of the convolution step at O(p·m), where m was the number of radial steps along each convolution ray sampled during the convolution. By convolving both directions at once, it reduced the number of function launches and increases performance. The result was written directly into the GPU's texture memory using a surface write functionality.

Summation. The NVB dose data resided in the GPU's texture memory after the surface write at the end of the line convolution kernel. A GPU thread was launched for every voxel from the

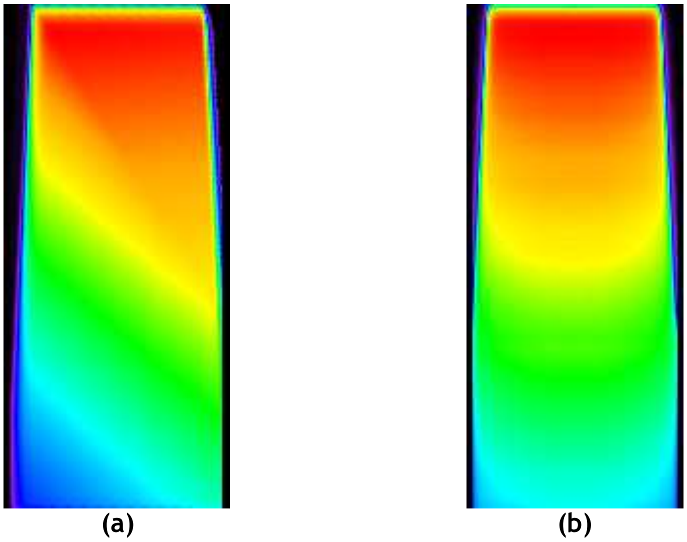


Figure 2.5. NVB Algorithm Summation. The convolved TERMA map for the current convolution direction is (a) converted back to the original voxelized coordinate system, and (b) summed with all other directions to obtain the final dose distribution.

original data set in the Cartesian coordinate system. The voxel's location in the convolution ray coordinate system was computed in order to sample the NVB dose data. The dose contribution of a single convolution direction converted back to Cartesian coordinates is shown in figure 2.5(a) by reading from the convolved dose that resided in texture memory. The intrinsic interpolation of texture memory was once again utilized to keep the computational complexity of this step to O(n), where n was the total number of voxels in the original data set. The final dose distribution was found by accumulating the contributions from each convolution direction, as shown in figure 2.5(b).

Quantifying GPU Convolution Accuracy and the Effect of the Sampling Parameters

To quantify the accuracy of the GPU implementation, we compared dose distributions for three digital phantoms with varying geometries, referred to hereafter as the accuracy phantoms, and a series of 12 segmented patient lungs. Shown in figure 2.5 are axial slices of the data sets used for the accuracy comparisons. Phantom A was a simple, homogenous block of water equivalent material. Phantom B, shown in figure 2.6(a), contained a cylinder and box of water equivalent

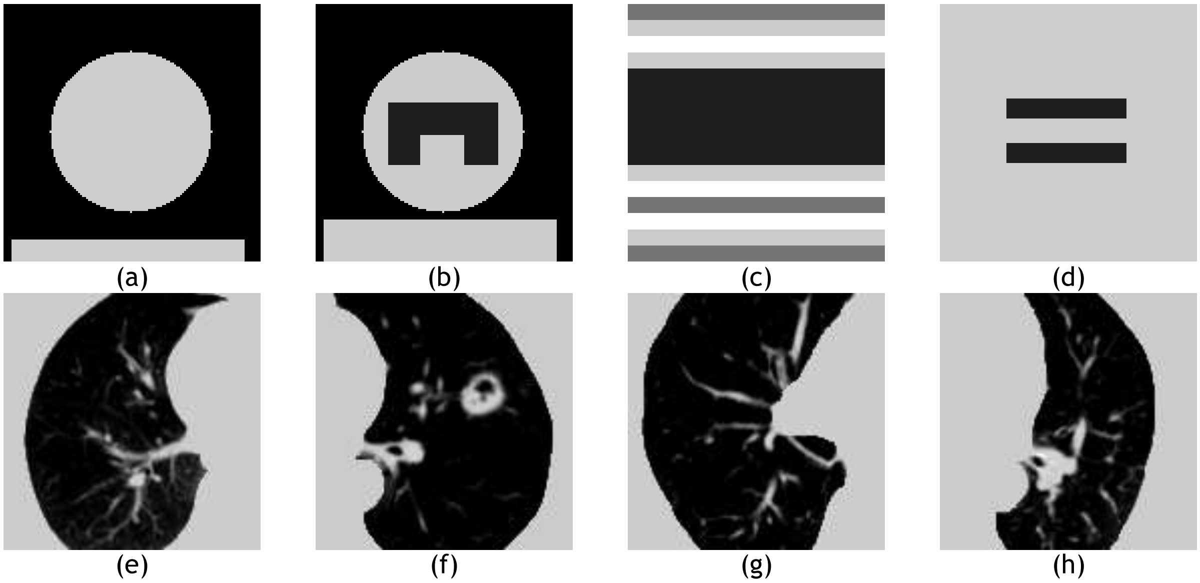


Figure 2.6. Density maps for the phantom data sets. Phantoms B, C (a-b), the classical slab phantom (c) and mediastinum phantom (d) were used for the dose accuracy studies. A sampling of the segmented patient lungs are also displayed (e-h). The segmented lungs were artificially surrounded with uniform water equivalent material. The homogenous water phantoms are not displayed due to simplicity.

material (density of 1 g/cm^3) surrounded by empty space/vacuum. Phantom C, shown in figure 1.6(b), introduced a lower density region (0.317 g/cm^3) within the cylinder, and serves as a simple lung phantom. The classical slab phantom and mediastinum phantom were also used for the accuracy study. The classical slab phantom, shown in figure 2.6(c), [REF - Ahnesjo] contained layers of adipose tissue, muscle, bone, and lung. The mediastinum phantom, shown in figure 2.6(d), has two low density boxes surrounded by water, simulating the lungs in the chest cavity. The resolution of the accuracy phantoms was isotropic 2 mm and the size of the matrix was 128x128x128.

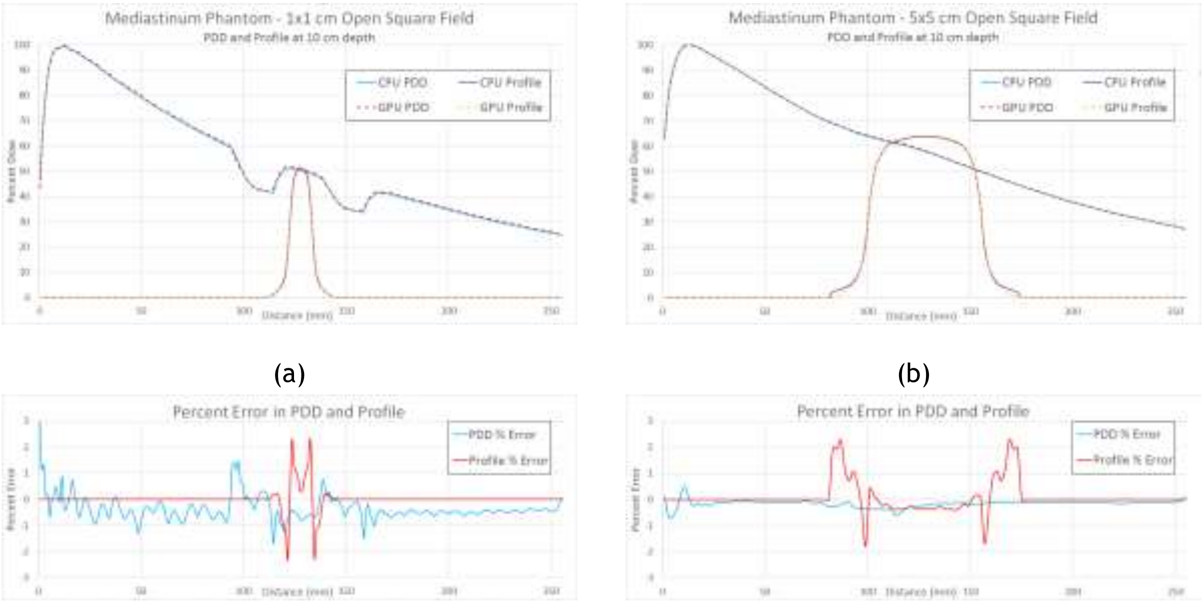
Real patient anatomy was also used for the accuracy calculations and a sample of the data sets are shown in figure 2.6(e-h). The lungs were segmented out and exported into 128 cube data blocks with voxels of in-plane resolution of 0.12 cm and slice thickness of 0.3 cm. The volume surrounding the lungs was set to have the density of 1 g/cm^3. The tested beam configuration was an open, square field whose isocenter was placed at the volumetric center of the data set. The spectrum was a discretization of a typical 6 MV treatment beam with a flattening filter.

The spectrum was divided into 14 mono-energetic bins. All dose distributions were evaluated using a three dimensional implementation of the gamma dose distribution comparison test[15, 26, 27], in addition to direct dose comparisons. The gamma value is the Euclidean distance between the reference dose distribution and the evaluated distribution. The gamma test has two test criteria; dose difference and distance to agreement, which were 2% and 0.2 cm, respectively. All gamma evaluations were performed on percent dose distributions, normalized to the maximum delivered dose. These criteria were well within clinical tolerances[28].

The NVB convolution method has four sampling parameters that can be optimized. The zenithal, azimuthal, and radial sampling of the original convolution, along with the parallel ray spacing introduced during ray-tracing. Setting the radial sampling and parallel ray spacing allowed the dose computation to be performed at a predetermined resolution, independent of the CT resolution. For the phantom studies, gamma results for voxels with zero density were ignored. The accuracy percentages represented the fraction of voxels within the volume of interest with accumulated dose that failed the gamma test. Ground truth was taken to be the CPU based calculation employing the highest number of zenithal and azimuthal sampling rates.

Performance Comparisons

To gauge the performance increases for the NVB GPU algorithm, we performed a series of tests on three homogenous water phantoms, hereafter referred to as the performance phantoms. Each performance phantom was a 256 mm cube, with isotropic resolutions of 1 mm, 2 mm, and 4 mm, which resulted in dimensions of 256^3, 128^3, and 64^3 voxels, respectively. The GPU algorithm was designed using NVIDIA’s CUDA, Compute Unified Device Architecture. GPU simulations were performed using an NVIDIA GTX 680 GPU, which has 1536 cores, and 2 GB of memory. The CPU was an Intel Core i7-3820 @ 3.60 GHz with 8 GB RAM. For inter-GPU comparison, a NVIDIA GTX 780 Ti GPU was employed, which has 2800 cores and 3 GB of memory.



(a) (b)

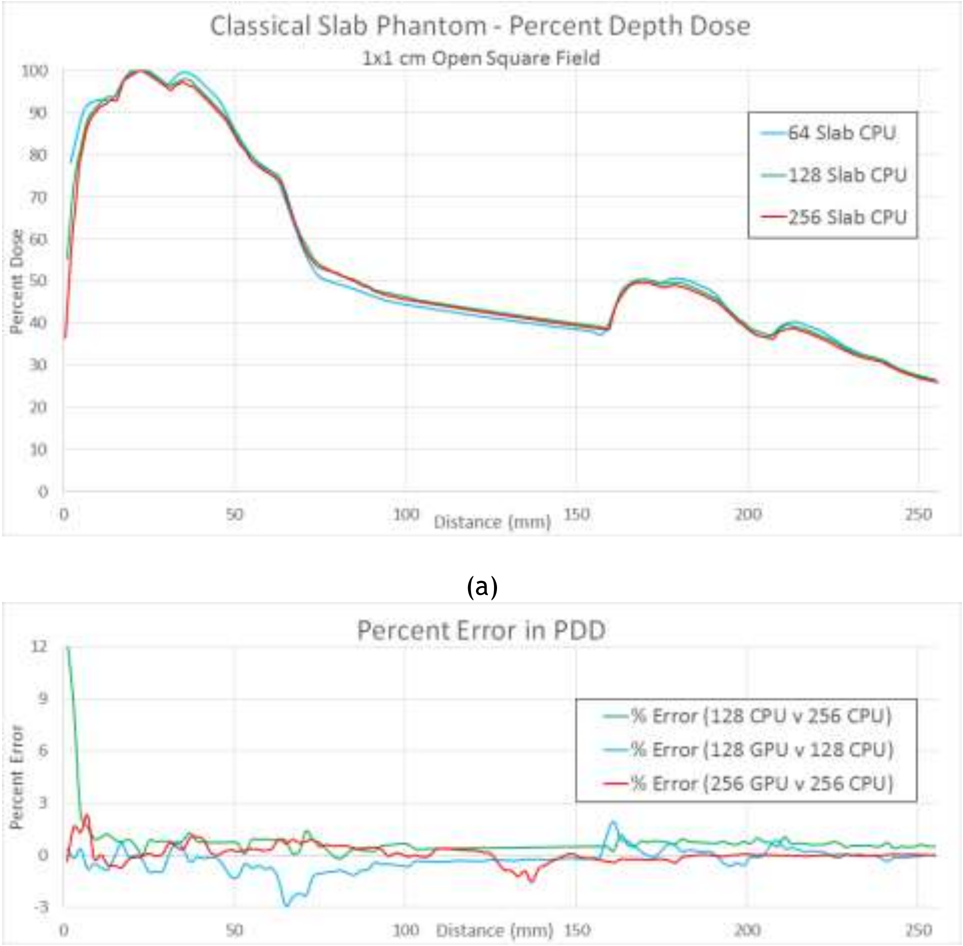
(c) (d)

Figure 2.7. Percent depth dose and cross profile comparisons. The PDD and profile for both the CPU convolution and the NVB GPU convolution for the mediastinum phantom irradiated with a 1x1 cm field is shown in (a). The corresponding percent error between the curves are displayed in (c). The same curves are displayed for the classical slab phantom irradiated with a 5x5 cm field in (b), with its respective percent error in (d).

RESULTS

GPU Accuracy

The accuracy of parallelizing the convolution algorithm was verified by examining the percent depth dose (PDD) and profile at 10 cm depth using direct dose comparison. Figure 2.7 displays the percent depth dose and cross profiles for the classical slab phantom and mediastinum phantom. Several beam sizes were examined, and the percent error was less than 1% for all voxels except for those with high dose gradients such as the penumbra and the build-up region. Much of the error seen between the CPU and GPU implementations can be attributed to the fact that the convolution is being calculated on different resolution grids. The NVB algorithm resamples the data according to the parallel ray-spacing and radial step size variables, and therefore is not convolving with exactly the same resolution as the voxel-based CPU algorithm. Figure 2.8(a) displays the results of convolving a 1x1 cm square field on the classical slab

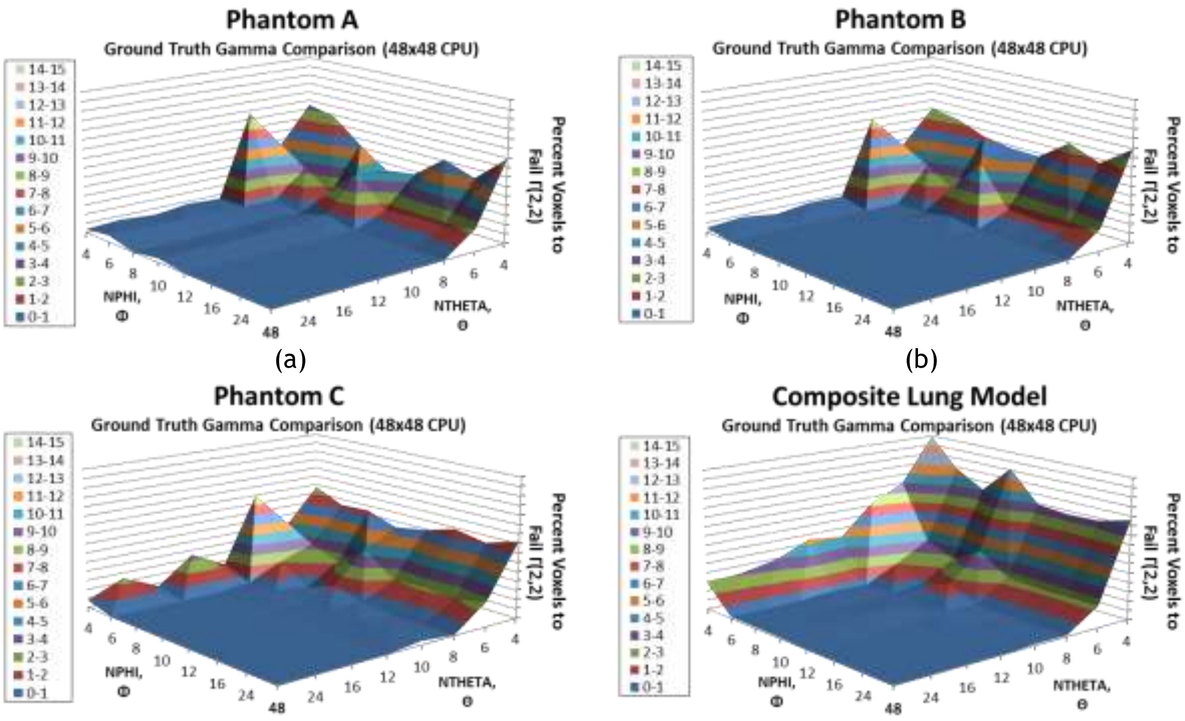


(a)

(b)

Figure 2.8. Percent dose difference for different resolution calculation grids during convolution. The percent depth dose curve for a 1x1 cm field on the classical slab phantom at three resolutions: 1, 2, and 4 mm isotropic voxels (a). (b) displays the percent dose difference between the 1 and 2 mm convolutions on the CPU, as well as the difference between CPU and NVB GPU convolutions when both are calculated at 1mm and 2mm.

phantom at three different resolutions. Figure 2.8(b) shows the percent error in the PDD between 1 mm and 2 mm resolutions of the CPU, compared to the error seen between the CPU and the NVB GPU algorithms. The error between the two CPU resolutions is on the same order as the error seen between CPU and GPU. The NVB algorithm was run using 1 mm radial step size and 1 mm parallel ray spacing. When the resolution of the CPU is comparable to the NVB coordinate system, the average error decreases from 0.26% to 0.15%. Again, the largest error is seen in the high dose gradient regions. Due to the intrinsic differences that arise from



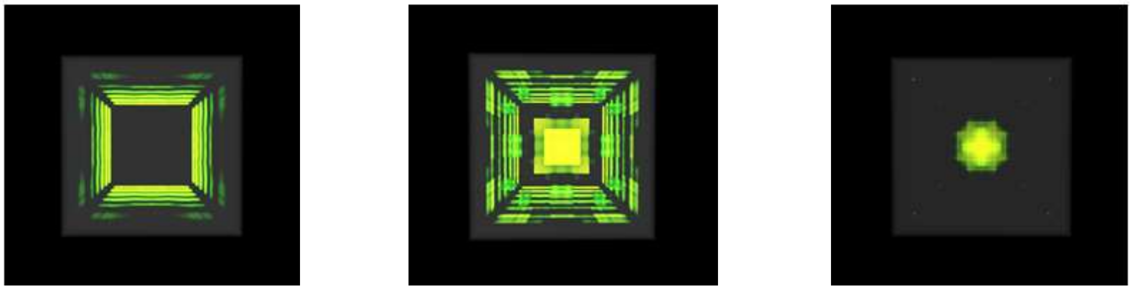
(a) (b)

(c) (d)

Figure 2.9. Dose accuracy as a function of angular sampling. The surface plots above display the percentage of voxels to fail a Gamma test with criteria of 2% and 2mm. The number of convolution rays in the zenithal and azimuthal directions were varied from 4 to 48. Ground truth was taken to be 48x48 rays. Plots (a), (b), and (c) display the plots for their respective phantoms, while plot (d) shows the failure percentage averaged over all twelve lung models.

convolution on different resolution calculation grids, we employed the 3D gamma dose distribution comparison tool when studying the effect of the sampling parameters on the accuracy of the NVB algorithm.

Performing gamma analyses over the entire spectrum of angular sampling combinations, using multiple field sizes and targets. When using the same sampling parameters as the ground truth calculations performed on the CPU, we observed that all voxels calculated using the non-voxel-based GPU parallelization passed the gamma test at 2% and 2mm. Such a result shows that the algorithm presented in this paper provided the same accuracy as that of clinically used dose convolution implementations.



(a) 48 zenithal angles x 4 azimuthal angles

(b) 4 zenithal angles x 4 azimuthal angles

(c) 4 zenithal angles x 48 azimuthal angles

Figure 2.10. Locality of dose discrepancies as a function of angular sampling. The figures display the results of a Gamma test with criteria of 2% and 2mm for a 128^3 phantom with 1 mm isotropic voxels irradiated with a 100x100 mm field at isocenter. (a) shows that reducing the azimuthal sampling causes errors in the penumbra region. (c) shows that reducing the zenithal sampling causes higher error along the beam axis. (b) shows the result for reduced sampling in both the zenithal and azimuthal directions.

Accuracy as a Function of Sampling Parameters

The plots in figure 2.9 display the percentage of voxels within the calculation cone that failed the gamma criteria as a function of angular sampling. Ground truth data were computed on the CPU using 48 zenithal and 48 azimuthal directions, with a 1 mm radial step size. Figure 2.9(a-c) shows the results for the accuracy phantom data sets. For an angular sampling combination of 8 zenithal and 8 azimuthal directions, the average failure percentage for the phantoms was just 0.012%, with a maximum of 0.082% for Phantom C. Figure 2.9(d) displays the average failure rates for the segmented lung data sets. The average failure rate was 0.74% for 8 zenithal and 8 azimuthal directions.

From the surface plots (figure 2.9), it is clear that for homogenous volumes such as phantom A, increasing the azimuthal sampling has little effect on the accuracy due to rotational symmetry about the beam direction. However, reducing the number of zenithal angles below 8 resulted in quickly increasing error because of the directionality of the kernel. They also show that for increasingly complex geometries, the total error became more dependent on the sampling rate. Particularly in the azimuthal direction, as shown when comparing phantom A, where the error due to azimuthal sampling was negligible, to the patient lung data sets.

Figure 2.10 displays a 3D rendering of the gamma results for Phantom A from a beam's eye point

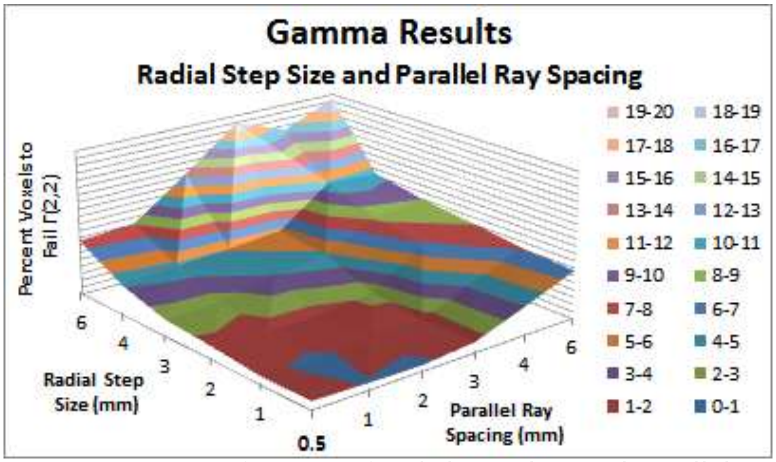
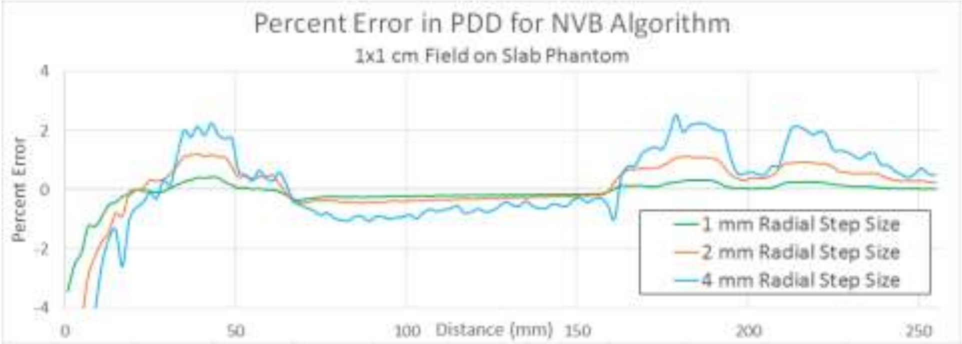


Figure 2.11. Dose accuracy as a function of radial sampling and parallel ray spacing. The surface plot displays the percentage of voxels to fail the Gamma test with criteria of 2% and 2mm. The radial sampling and parallel ray spacing define the new non-voxel coordinate system. The size of the sampling steps has a drastic effect on the accuracy of the dose calculation.

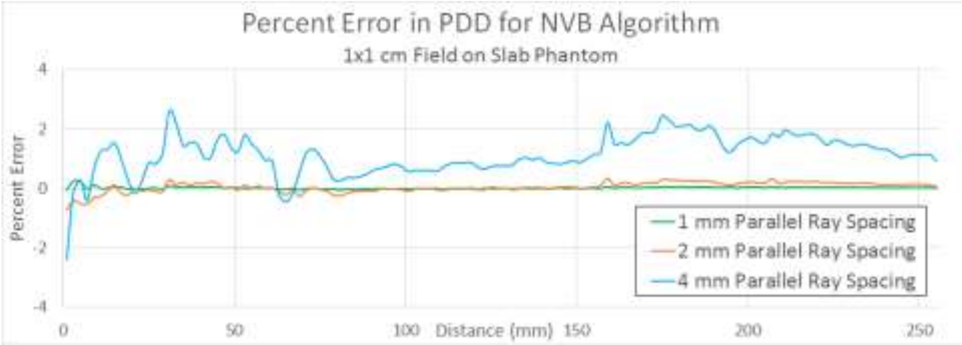
of view. The volume itself can be seen as a gray cube, while the failing voxels are overlaid with a heat map. Three angular sampling combinations are displayed, illustrating the effects of reduced sampling in both the zenithal and azimuthal directions. Reducing the azimuthal sampling increased the discrepancies in the penumbra regions of the beam, while reducing the zenithal sampling causes more significant deviations along the beam axis. Figure 2.11 plots the percent of voxels to fail a gamma test at 2% and 2 mm when increasing the radial sampling and parallel ray spacing. Increasing either the radial step size or the parallel ray spacing any higher than 2 mm caused rapidly increasing dose distribution modeling errors. The best results were seen when the radial step size was the same as the ground truth at 1mm, and the parallel ray spacing was less than or equal to 2mm.

To further illustrate the influence of the parallel ray spacing and radial step size on the integrity of the dose calculation, figure 2.12 displays the percent dose difference in the PDD for a 1x1 cm field on the classical slab phantom.

The plots in figure 2.13 show the percent dose difference for the depth profile along the central beam axis and the beam profile perpendicular to the beam through isocenter for different sets



(a)

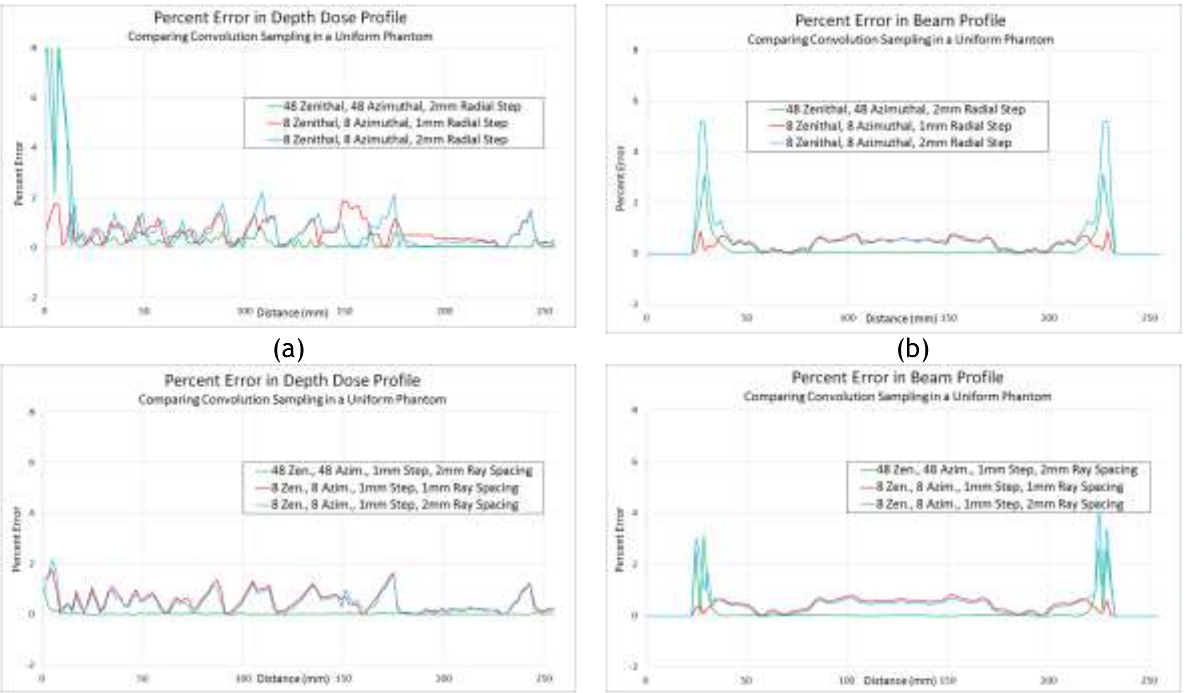


(b)

Figure 2.12. Effect of parallel ray spacing and radial step size on the percent dose difference as a function of depth. In these experiments, the classical slab phantom was irradiated with a 1x1 cm square field. (a) shows the effect of increasing the radial step size, where 0.5 mm is taken as ground truth and all other parameters are held constant. (b) shows the effect of increasing the parallel ray spacing, where again 0.5 mm is taken as ground truth and all other parameters are held constant.

of sampling parameters. Comparing the percent dose difference between the 48x48 and the 8x8 angular sampling combination, the error was less than 1% for the majority of the depth profile and beam profile, and the maximum error was less than 2%. This bolstered the conclusion that the 8x8 angular sampling combination provided clinically acceptable accuracy, even without considering the distance to agreement criterion of the gamma distribution analysis method. Figures 2.13(a-b) illustrate how reducing the radial sampling caused large errors in the build-up region and penumbra, even when the increase was as small as 1 to 2 mm.

Figures 2.13(c-d) display similar plots where radial sampling was constant at 1 mm and the parallel ray spacing was varied between 1 and 2 mm. As shown by the two lines with 8x8 angular sampling, increasing the parallel ray spacing caused little to no increase in the error.



(a) (b)

(c) (d)

Figure 2.13. Percent difference in dose profiles due to sampling. Plot (a) shows the percent difference from ground truth for the depth profile along the beam axis for three combinations of zenithal, azimuthal, and radial sampling. (b) shows the percent difference along the beam profile through isocenter for the same sampling combinations. Plots (c) and (d) introduce the effect of the parallel ray spacing during the ray-tracing step in the non-voxel based algorithm.

GPU Performance

Table 2.1 gives the computation time for the generic GPU implementation and the non-voxel-based implementation fully optimized for the GPU architecture, calculated on three homogenous water performance phantoms. This computation times reported for both the CPU and the GPU encompasses only the convolution calculation. For the GPU this includes all kernel calls (4 per convolution direction), and for the CPU this includes all calculations within the outer-most loop of the convolution. The TERMA was calculated previously and already resided in the global memory of the GPU. The average time of the TERMA calculation was 1 ms. Similarly, the density matrix was also residing in the GPU's global memory, so the extra overhead due to memory copies from CPU to GPU was minimal. On average, including the TERMA computation and memory copies added 5-6 ms to the overall computation time. The times are displayed for combinations of 24 zenithal angles with 16 azimuthal angles for 384 total rays

Table 2.1. Computation Times

	Generic GPU	NVB GPU	Generic GPU	NVB GPU
GPU Hardware	GTX 680	GTX 680	GTX 680	GTX 680
Directions	384 rays	384 rays	64 rays	64 rays
256x256x256 (1mm resolution)	45.03 s	2.04 s	7.42 s	0.343 s
128x128x128 (2 mm resolution)	8.01 s	1.61 s	1.30 s	0.274 s
64x64x64 (4 mm resolution)	2.50 s	1.70 s	0.42 s	0.282 s

(commonly used parameters for comparison testing [14], and 8 zenithal angles with 8 azimuthal angles for 64 total rays, using a 100mm square field at isocenter. The computation times for both the generic GPU and NVB algorithms were linearly related to the number of convolution rays. For the most computationally expensive calculation, the NVB algorithm improved the calculation time from 45 seconds to 2 seconds, a speed factor increase of over 22. The accuracy study presented above showed that an angular sampling combination of 8x8 produced acceptable results, and the total convolution time for the highest resolution phantom, a 256 voxel cube with 1 mm isotropic voxels was less than 350 ms for the non-voxel-based algorithm.

Table 2.2. Average Acceleration Using the GPU Parallelization

Acceleration Phantom Resolution	64^3 Phantom 4 mm voxels	128^3 Phantom 2 mm voxels	256^3 Phantom 1 mm voxels
CPU / Generic GPU	59.26 ± 1.66	113.2 ± 1.75	193.7 ± 12.7
Generic GPU/ NVB GPU	1.46 ± 0.04	4.8 ± 0.14	21.6 ± 0.6
CPU / NVB GPU	86.63 ± 3.49	546.4 ± 20.3	4,175.5 ± 354.9

Table 2.2 presents the performance gains when comparing identical sampling parameters across all three algorithms. The results presented were averaged over every angular sampling combination, and are shown with the standard deviation. For the 64^3 phantom, the generic GPU parallelization technique provided an acceleration factor of nearly 60 over the CPU. The NVB implementation boosted the performance to more than 86 times over the CPU. The comparison showed the NVB implementation increased performance 1.46x over the generic GPU implementation on average. This advantage grew as the data size and computational complexity increased.

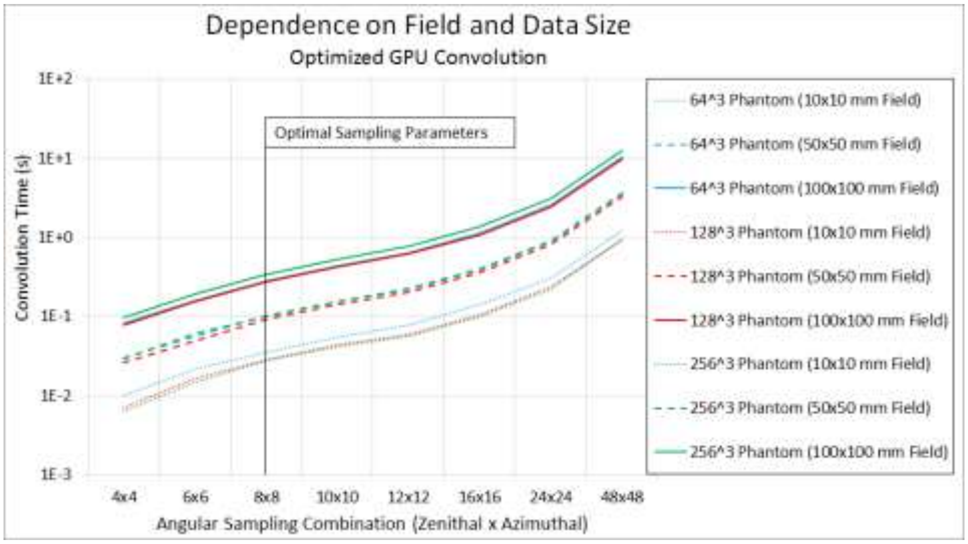


Figure 2.14. Convolution computation time as a function of field size and data size. The plot displays the convolution time as a function of the angular sampling combination for three different field sizes on each of the performance phantoms. The results are clearly grouped by field size, while there is little distinction between the phantom data size.

As seen from the ratios of the CPU over the generic GPU convolution times, the advantage of the GPU increased with an increase in the data resolution because the resolution was directly related to computational effort. However, the ratios of the generic GPU implementation convolution time over the NVB implementation convolution time showed that the optimized memory accesses of the NVB method were able to maintain significantly higher throughput as computational effort increased. The NVB method was nearly 22 times faster than the generic GPU implementation for the 256^3 phantom. This resulted in a total acceleration factor of more than 4000 when comparing the CPU algorithm against the NVB GPU parallelization technique. The significant improvement from the generic GPU method to the NVB method could be attributed to an intrinsic data size independence of the NVB technique, which will be further discussed in the next section.

Field Size and Data Size Dependence

An advantage of transforming the convolution into a non-voxel-based coordinate system was

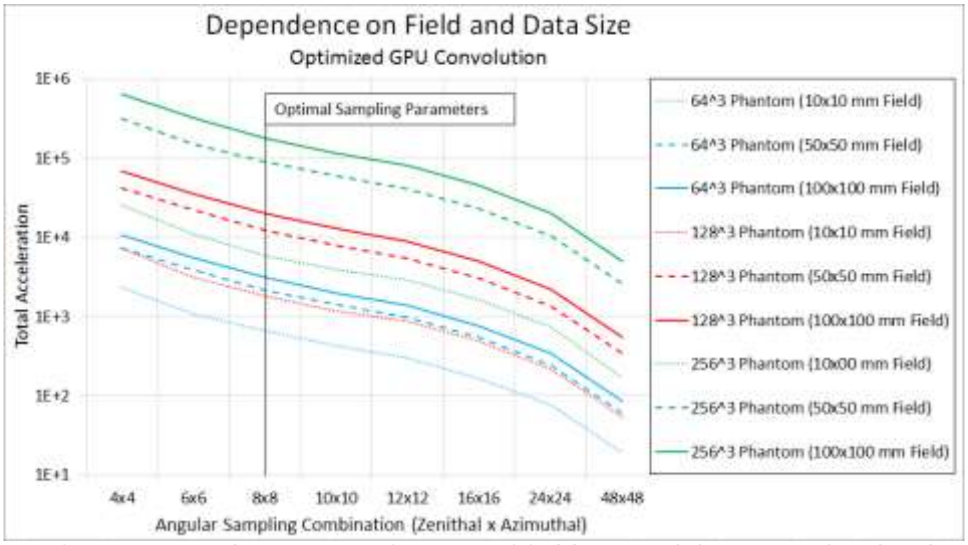


Figure 2.15. Performance acceleration as a function of field size and data size. The plot shows the total acceleration due to parallelization on the GPU and reduced angular sampling. Greater computational effort results in greater acceleration. The highest resolution phantom irradiated with the largest field size results in the most voxels involved in the convolution calculation and the greatest acceleration. The data size has a strong correlation with acceleration, as the 10x10 mm^2 field size on the 256^3 phantom shows greater acceleration than the 100x100 mm^2 field size on the 64^3 phantom.

that the calculation grid was then controlled exclusively by the sampling parameters. The resolution of the grid in the non-voxel-based system was determined by the radial step size and the parallel ray spacing, and the number of rays cast through the volume depended only on the size of the field and the parallel ray spacing. This eliminated the dependence of the computation time on the original data resolution. Figure 2.14 shows the convolution time for the non-voxel-based algorithm using a 10x10 mm^2 field, a 50x50 mm^2 field, and a 100x100 mm^2 field, over a spectrum of angular sampling combinations. The data is clearly grouped by field size, but more interestingly is the lack of separation across the data size. The non-voxel-based algorithm proved to be independent of the data volume because it resampled the data according to the radial step size and the parallel ray spacing parameters. This caused large performance gains over the CPU algorithm.

Sampling Acceleration

The ground truth calculation time was taken as the maximum convolution sampling combination of 48 zenithal and azimuthal angles. The sampling acceleration was directly related to the number of rays used during convolution. Reducing the angular sampling of each angle by a factor of 2 resulted in 4 times speed up, and so forth. By reducing the angular sampling to 8 zenithal angles and 8 azimuthal angles, the performance was increased by a factor of 36. As shown in figure 2.15, combining the reduced sampling acceleration with the acceleration provided from the non-voxel-based GPU algorithm pushed the maximum acceleration over 175,000 times for the 256 voxel cube phantom and a 100x100 mm^2 field. While the convolution times were very similar across data sizes for the NVB algorithm, figure 2.15 shows a fairly consistent increase in acceleration around one order of magnitude as the data size increased. Also, the smallest field on the highest resolution phantom saw larger accelerations than the largest field on the lowest resolution phantom. The total combined acceleration for both the generic GPU and NVB implementations from GPU parallelization and reduced sampling are presented in table 3 for each of the three acceleration phantoms.

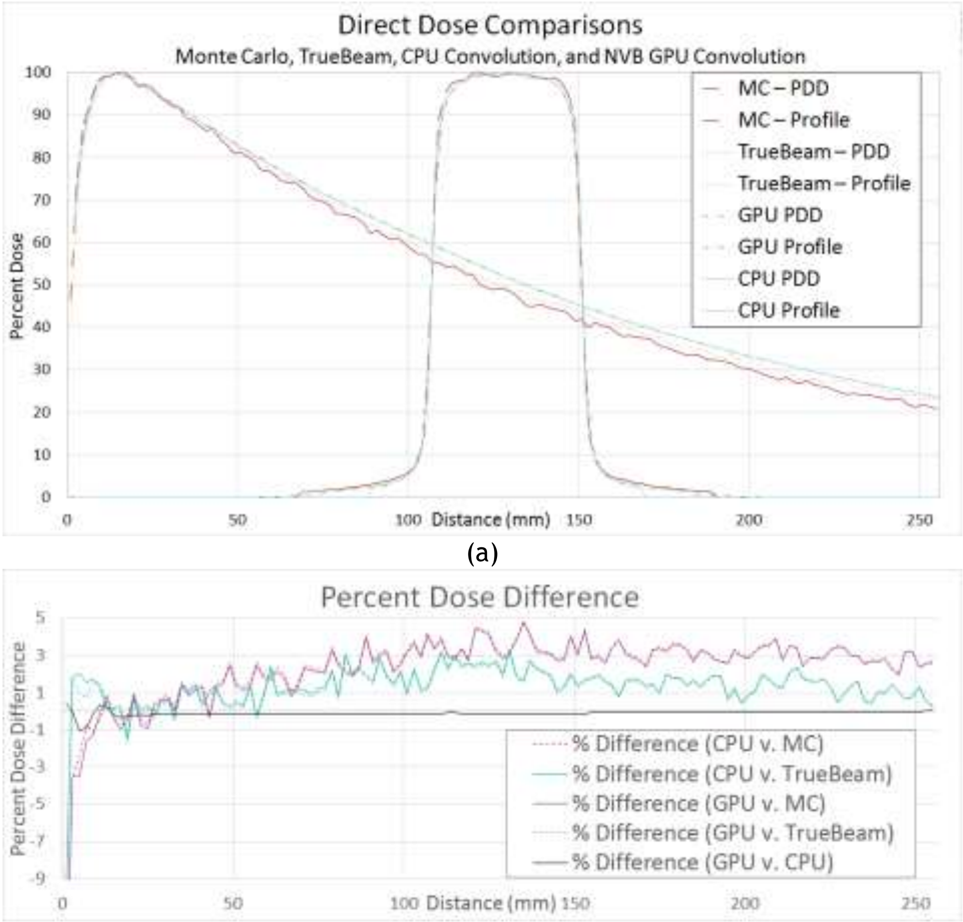
Table 2.3. Acceleration Using the Optimal Sampling Parameters and GPU Parallelization

	64^3 Phantom 4 mm voxels	128^3 Phantom 2 mm voxels	256^3 Phantom 1 mm voxels
CPU / Generic GPU	2,100	4,100	8,200
CPU / NVB GPU	3,100	19,500	176,000

DISCUSSION

The convolution is scalable to multiple GPUs. Theoretically, there was a direct relationship between the number of GPUs employed and the performance gains for large workloads[29]. These simulations used an open square field for verification and comparison. Utilizing multiple GPUs would allow calculating treatment plans with multiple fields. Incorporating complex beam geometries, varied fluence maps as are generated for IMRT, multiple fields, and Multi-Leaf Collimators are all aspects that can be investigated.

Figure 2.16(a) shows the depth dose and cross profiles for a 5x5 cm field in a homogenous water phantom from a Monte Carlo generated dose distribution using the same energy spectrum as



(a)

(b)

Figure 2.16. Direct dose comparisons between Monte Carlo, TrueBeam, voxel-based CPU convolution, and NVB GPU convolution. The depth dose and cross profiles for a 5x5 cm square field at 10 cm depth are shown for each dose distribution source in (a). The percent difference for the depth dose profiles are shown in (b).

the convolution algorithms, a Monte Carlo generated distribution using the phase space energy spectrum data from a Varian TrueBeam® linear accelerator with flattening filter provided by Varian Medical Systems, the voxel-based CPU convolution used as ground truth in this paper, and the NVB GPU convolution presented in this paper. The Monte Carlo generated dose distributions were created using MCNP4C, with histories of $2x10^9$ photons to achieve less than 2% statistical variation. Both the CPU convolution and the NVB convolution were calculated using 8 azimuthal angles and 8 elevation angles, with a radial step size of 1 mm. Additionally, the NVB convolution used a parallel ray-spacing of 1 mm. Significant differences can be seen along each profile between the convolution methods and the other data. Figure 2.16(b) plots the percent dose difference for both the CPU convolution and the NVB GPU convolution. While the CPU convolution is regarding as ground truth in this paper, the plot shows that there is definitely room for improvement to more realistically recreate the actual dose distributions measured from the treatment machine, and the gold standard Monte Carlo dose calculations. With further performance enhancement, we should be able to deconstruct the poly-energetic kernel and calculate the energies independently, which will eliminate some assumptions and estimations currently used in the convolution/superposition algorithm and produce a dose distribution closer to the Monte Carlo distribution.

Table 2.4 tabulates the computation time dedicated to each of the four components of the non-voxel based algorithm for both the high and low convolution ray count and each of the performance phantoms. The performance bottleneck still resided with the convolution step, due to the requirement for ray-casting along the line to accumulate the effective radiological distance when applying the heterogeneity correction.

Table 2.4. Percentage of Total GPU Computation Time for NVB Convolution Algorithm

Acceleration Phantom	64^3 Phantom		128^3 Phantom		256^3 Phantom	
Directions	384 rays	64 rays	384 rays	64 rays	384 rays	64 rays
Ray-tracing	4.51	4.40	10.95	10.68	19.45	19.67
Transposition	4.60	4.50	3.96	3.92	2.96	2.96
Line Convolution	90.06	90.27	82.17	82.54	61.10	61.02
Summation	0.83	0.83	2.92	2.85	16.49	16.35

We are currently investigating a method to stretch the data volume according to effective radiological distance during the initial ray-tracing. The convolution could then be even further parallelized as each data point in the non-voxel based coordinate system would represent an equal amount of attenuation. The algorithm would no longer have to step along the rays to apply the heterogeneity correction, but simply apply multiplication and summation reduction techniques which are much more suitable for parallel architecture.

As the computing hardware continually improves, the software design considerations discussed in this paper become more and more critical to maximizing performance. Just as table 2.1

reported the improvement in computation time of the NVB algorithm compared to the generic GPU algorithm, table 2.5 compares the performance of our NVB algorithm for the GTX 680, which the algorithm was developed on, and the newest card released by NVIDIA, the 780 TI. An average speed up over 1.8 times was seen for both high and low number of convolution rays on all three of the performance phantoms.

Table 2.5. NVB Computation Times with Improving Hardware

Directions	384 rays		64 rays	
GPU Hardware	GTX 680	GTX 780 TI	GTX 680	GTX 780 TI
256x256x256	2.04 s	1.06 s	0.343 s	0.177 s
128x128x128	1.61 s	0.89 s	0.274 s	0.149 s
64x64x64	1.70 s	0.98 s	0.282 s	0.161 s

CONCLUSIONS

The convolution parameters (zenithal angle sampling, azimuthal angle sampling, radial step size, and parallel ray spacing) could be optimized for maximum acceleration with minimal loss of accuracy. This was demonstrated by performing dose calculations using five digital phantoms and twelve patient lung CTs. In both cases, a zenithal/azimuthal combination of 8/8 provided the best performance while maintaining accuracy. Both the phantoms and lung models passed a gamma test of 2% or 2mm at better than 99%.

Splitting the acceleration between the sampling optimization and GPU implementation showed a consistent speed up of about 36 when reducing the convolution sampling from 48/48 to 8/8, while the GPU implementation provided a second improvement level between 86 and nearly 4200 times speed up depending on the data size and resolution. This resulted in total performance gains of just over 3000 times for the smallest 64 voxel performance phantom and over 175,000 times for the largest 256 voxel performance phantom when compared to a non-optimized CPU algorithm. Optimizing the NVB algorithm for the GPU architecture also improved the performance significantly compared to a generic GPU implementation, providing nearly 22 times speed up for the 256 voxel performance phantom.

Future work will see the expansion of our non-voxel-based convolution to a multi-GPU framework. Implementing the outlined optimization strategies and eliminating many of the assumptions and estimations currently employed by convolution/superposition to reduce computation times, this method can improve both accuracy and speed for computing on-the-fly dose distributions. These improvements are valuable for the clinical dosimetry efficiency and will facilitate real-time adaptive radiotherapy.

REFERENCES

[1] Britton, K., Dong, L., and Mohan, R., "Image Guidance to Account for Interfractional and Intrafractioin Variations: From a Clinical and Physics Perspective," in [Image-Guided Radiotherapy of Lung Cancer], J. Cox, J. Chang, and R. Komaki, Editors, Informa Healthcare USA, Inc.: New York, NY (2007).

[2] Li, X. A., Liu, F., Tai, A., and Ahunbay, E., "Development of an online adaptive solution to account for inter- and intra-fractional variations," Radiotherapy and Oncology 100(3), 370-374 (2011).

[3] Liu, Erickson, Peng, and Li, "Characterization and Management of Interfractional Anatomic Changes for Pancreatic Cancer Radiotherapy," International Journal of Radiation Oncology, Biology, Physics 83(3), e423-e429 (2012).

[4] Stewart, J., Lim, K., Kelly, V., and Xie, J., "Automated Weekly Replanning for Intensity-Modulated Radiotherapy of Cervix Cancer," International Journal of Radiation Oncology, Biology, Physics 78(2), 350-358 (2010).

[5] Bujold, A., Craig, T., Jaffray, D., and Dawson, L., "Image-Guided Radiotherapy: Has It Influenced Patient Outcomes?," Seminars in Radiation Oncology 22(1), 50-61 (2012).

[6] Xing, L., Siebers, J., and Keall, P., "Computational Challenges for Image-Guided Radiation Therapy: Framework and Current Research," Seminars in Radiation Oncology 17(4),

245-257 (2007).

[7] Sanders, J. and Kandrot, E., [CUDA By Example], Boston, MA: Pearson Education, Inc., (2011).

[8] Kirk, D. and Hwu, W.-m., [Programming Massively Parallel Processors], Burlington, MA: Elsevier, Inc., (2010).

[9] Riha, L. and El-Sayed, H., *Real-Time Motion Object Tracking Using GPU*, in *International Conference on Computer Systems and Applications*. 2011. p. 301-304.

[10] Hissoiny, S., Ozell, B., and Despres, P., "Fast convolution-superposition dose calculation on graphics hardware," Medical Physics 36(6), 1998-2005 (2009).

[11] Hissoiny, S., Ozell, B., and Despres, P., "A convolution-superposition dose calculation engine for GPUs," Medical Physics 37(3), 1029-1037 (2010).

[12] Jacques, R., Taylor, R., Wong, J., and McNutt, T., "Towards real-time radiation therapy: GPU accelerated superposition/convolution," Computer Methods and Programs in Biomedicine (2009).

[13] Jacques, R., Wong, J., Taylor, R., and McNutt, T., "Real-time dose computation: GPU-accelerated source modeling and superposition/convolution," Medical Physics 38(1), 294-305 (2010).

[14] Chen, Q., Chen, M., and Lu, W., "Ultrafast convolution/superposition using tabulated and exponential kernels on GPU," Medical Physics 38(3), 1150-1161 (2011).

[15] Chen, Q. and Lu, W., "Validation of GPU based TomoTherapy dose calculation engine," Medical Physics 39(4), 1877-1886 (2012).

[16] Lu, W., "A non-voxel-based broad-beam (NVBB) framework for IMRTtreatment planning," Physics in Medicine and Biology 55, 7175-7210 (2010).

[17] Mackie, "A convolution method of calculating dose for 15-MV x-rays," Medical Physics 12, 188-196 (1985).

[18] Ahnesjo, A., "Collapsed cone convolution of radiant energy for photon ose calculation in heterogeneous media," Medical Physics 16, 577-592 (1989).

[19] Siddon, R., "Fast calculation of the exact radiological for a three-dimensional array," Medical Physics 12(2), 252-255 (1985).

[20] Hoban, P. W., "Accounting for the variation in collision kerma-to-terma ratio in polyenergetic photon beam convolution," Medical Physics 22(12), 2035-2044 (1995).

[21] Hubbell, J. H. and Seltzer, S. M. *Tables of X-Ray Mass Attenuation Coefficients and Mass Energy-Absorption Coefficients from 1 keV to 20 MeV for Elements Z = 1 to 92 and 48 Additional Substances of Dosimetric Interest*. 2011; Available from: http://www.nist.gov/pml/data/xraycoef/index.cfm.

[22] Mackie, T. R., Bielajew, A. F., Rogers, D. W., and Battista, J. J., "Generation of photon energy deposition kernels using the EGS Monte Carlo code," Physics in Medicine and Biology 33(1), 1-20 (1988).

[23] Lu, W., Olivera, G., Chen, M.-L., Reckwerdt, P., and Mackie, T., "Accurate convolution/superposition for multi-resolution dose calculation using cumulative tabulated kernels," Physics in Medicine and Biology 50(4), 655-680 (2005).

[24] Hoban, P. W., Murray, D. C., and Round, W. H., "Photon beam convolution using polyenergetic energy deposition kernels," Physics in Medicine and Biology 39(4), 669-685 (1994).

[25] Qin, B., Wu, Z., Su, F., and Pang, T., [GPU-Based Parallelization Algorithm for 2D Line Integral Convolution], Lecture Notes in Computer Science, Vol, 6145: Springer, (2010).

[26] Gu, X., Jia, X., and Jiang, S., "GPU-based fast gamma index calculation," Physics in Medicine and Biology 56(5), 1431-1441 (2011).

[27] Low, D. and Dempsey, J., "Evaluation of the gamma dose distribution comparison method," Medical Physics 30(9), 2455-2465 (2003).

[28] Fraass, B., Doppke, K., Hunt, M., Kitcher, G., Starkschall, G., Stern, R., and Dyke, J. V., "American Association of Physicists in Medicine Radiation Therapy Committee Task Group 53: Quality Assurance for clinical radiotherapy treatment planning," Medical Physics 25(10), 1773-1829 (1998).

[29] Santhanam AP, Y, M., H, N., N, P., SL, M., and Kupelian, P., "A multi-GPU real-time dose simulation software framework for lung radiotherapy," International Journal of Computer Assisted Radiology and Surgery 7(5), 705-719 (2011).

CHAPTER 3: Analytical modeling and implementation of a multi-GPU cloud-based server (MGCS) framework for non-voxel-based dose calculations

A version of this chapter has been submitted for publication as a manuscript to the International Journal of Computer Assisted Radiology and Surgery

ABSTRACT

Purpose. In this paper, a multi-GPU cloud-based server (MGCS) framework is presented for dose calculations, exploring the feasibility of remote computing power for radiotherapy purposes. An analytical model was developed to describe potential performance in the MGCS environment, in order to intelligently determine the workload distribution and theoretical limit of acceleration.

Methods. Numerical studies were performed using a cloud-based computing setup that consisted of 14 NVidia GPUs distributed over 4 server nodes interconnected by a 1 Gigabits per second (Gbps) network. Inter-process communication methods among the multi-GPU processes were optimized to both facilitate the distribution of computing resources as well as compress and minimize data transfers over the server interconnect.

Results. The computation time predicted by the analytical model matched experimentally observed computation times within 1-5%. The theoretical limit of acceleration for an MGCS implementation compared to a local, single-GPU implementation was directly proportional to the total number of GPU devices utilized. The MGCS performance approached this limit when the computational tasks far outweighed the memory operations. The multi-GPU cloud server implementation reproduced the results of the original NVB dose computation with negligible numerical differences from distributing the work among several processes and the implemented data reduction strategies.

Conclusions. The results showed that a cloud-based computation engine was a feasible solution for enabling clinics to make use of fast dose calculations for advanced treatment planning and adaptive radiotherapy. The cloud-based system was able to exceed the performance of a local machine even for optimized, simple calculations, and provided significant acceleration for computationally

intensive tasks. Such a framework can provide access to advanced technology and computational methods to many clinics, providing an avenue for standardization across institutions without the requirements of purchasing, maintaining, and continually updating hardware.

INTRODUCTION

Radiotherapy has been an effective tool in treating many types of cancers, and with recent advancements in dose conformity and improved tumor targeting, significant improvements in treatment efficacy have been observed [1-6]. Adaptive radiotherapy (ART) is one such critical tool in further improving radiotherapy treatments. Undetected and uncompensated changes in the patient anatomy may lead to a reduction in treatment efficacy and subsequently in the patient's quality of life. ART has enabled treatment plans to be altered to compensate for changes in patient anatomy due to setup variation, physiological regression, and radiation response [7]. One of the critical components required for broad on-table ART implementation is real-time dose calculation in order to allow the treatment plan to be updated in the time between the daily positioning imaging and beam on without reducing the clinical throughput. Factors that increase computational complexity of re-planning in real-time include (a) complex treatment delivery systems [8, 9], (b) treatment plans that aggressively reduce dose to the surrounding organs [7], and (c) complex physiological regression in the patient anatomy [10]. Increasing the computational speed-up of the dose-convolution will be a critical component to enabling on-table re-planning for ART implementation. The difficulty for a clinic to take full advantage of advancing technology arise from the space required to house the hardware and the necessity to continually upgrade it. Cloud based computing can give access to far more computational power than would be feasible locally, while constantly expanding and rotating in new hardware as it becomes available.

Recent improvements in near real-time dose convolutions stem from a non-voxel based (NVB) dose convolution approach and its parallelized implementation [11]. Accelerated dose computations can now approach real-time, but require advanced hardware, such as graphics processing units (GPU). General purpose GPU computing is a rapidly developing field, with new generations released more than once per year. Expanding the calculation to a multi-GPU implementation further improves performance in a linear relationship to the number of GPUs employed [12]. To take full advantage of the potential processing power would require constant hardware updates. Extending these algorithms to run on cloud-based GPU servers would allow clinics to fully utilize these methods without the direct cost or overhead of installing and continuously updating computing hardware. Recent cloud computing works have explored the performance benefits for Monte Carlo based dose simulations [13-16]. Additionally, Meng et al. explored utilizing Google's MapReduce technology for scaling CT reconstruction using cloud computing technology [17].

In 2013, Kagadis et al. presented a thorough introduction to cloud computing, with a review of the recent medical imaging applications [18]. Moore et al. followed this paper in 2014 with a review of advanced computing methods in radiation oncology, and a discussion on promising developments for the near future [19]. These works describe the potential advantages of a cloud computing environment, including access to more extensive computing power and storage, removing the responsibility of maintaining and updating the computer hardware from the clinical institution, and the possibility for shared information between institutions and promoting standardization and collaboration between clinics [20-22]. However, the application of such frameworks utilizing GPUs for radiation therapy purposes remains largely unexplored [23].

In this paper, we present a cloud-based GPU-accelerated dose calculation engine and an analytical model to describe the potential performance of the cloud-based implementation as a function of the distribution of tasks. Specifically, we investigate the feasibility of performing a NVB dose convolution in a multi-GPU cloud-based setup, which provides an additional layer of parallelization for an algorithm already optimized for GPU architecture. A cloud-based system consisting of a set of 12 NVidia GTX 980 GPUs and 2 GTX 780 Ti GPUs were employed for this study. The GPUs were distributed in 4 different server nodes interconnected by a 1 Gbps interconnect. Key contributions of the method include (a) optimizing inter-process communication (IPC) between and within cloud server nodes to reduce latencies of memory operations, (b) developing a model to predict performance and acceleration against a single-threaded, single-GPU implementation, and (c) evaluating the feasibility of utilizing a multi-GPU cloud-based server framework for general radiotherapy tasks.

MATERIALS and METHODS

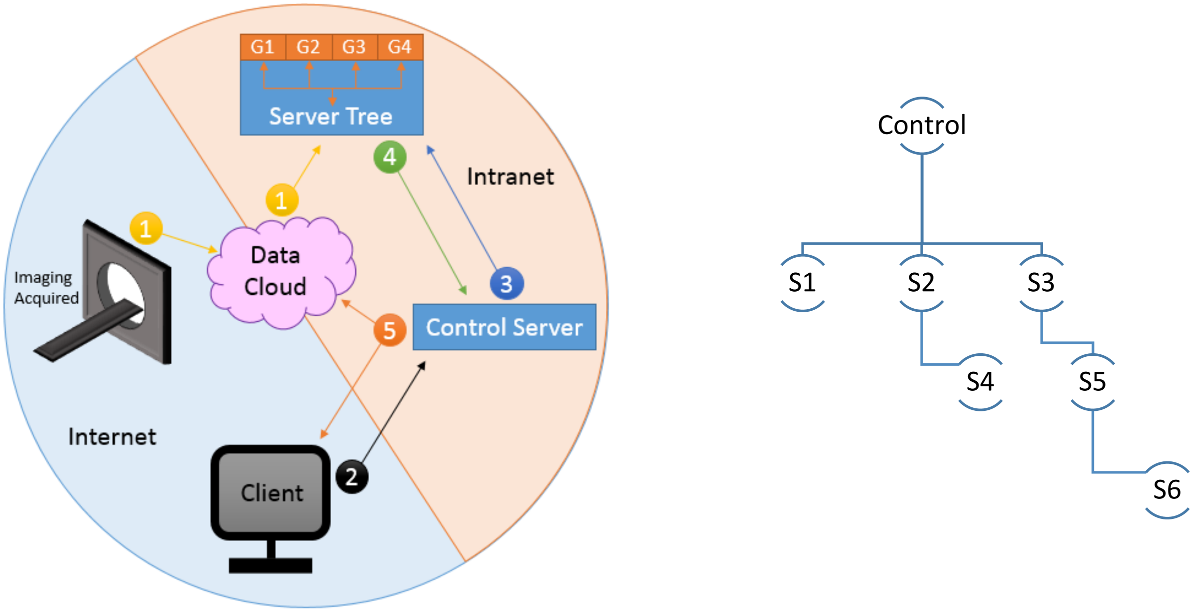
Non-Voxel-Based (NVB) Convolution Implementation

A full description of the NVB dose calculation framework was published in Neylon et al [11]. However, for clarity, a brief description is presented. The NVB algorithm was developed to optimize the convolution/superposition algorithm for GPU architecture, focusing on maximizing the efficiency of GPU memory usage and access patterns. The algorithm looped over the convolution rays, and converted the data volumes to a NVB coordinate system by ray-tracing through the volumes along each convolution direction.

Cloud-Based Multi-GPU Framework Approach

Extending the NVB algorithm to multi-GPU. The NVB dose convolution algorithm lent itself well to a multi-GPU implementation. Each convolution direction was already being calculated independently and summed afterwards to obtain the final dose distribution. Thus, the GPU convolution for a single direction could be extracted as a separate process, and the convolution directions were distributed among the available GPU devices. The results from all convolution directions can be consolidated using IPC, then summed and normalized to produce the final dose distribution. The data required for the convolution were pre-computed and made available to all processes.

Cloud-based dose computation workflow. Figure 3.1(a) provides a schematic representation of the possible workflow using the MGCS framework for dose computations. During the treatment planning stage, a patient CT is acquired. The imaging data is uploaded to the cloud storage and synced with the individual server nodes (step 1). The role of cloud storage has been previously exemplified by



(a) Multi-GPU Cloud Schematic (b) Branching Server Tree

Figure 3.1. (a) Graphical schematic of MGCS pipeline. (1) Imaging acquired, pushed to cloud and synced to server node tree. (2) Remote client signals the control server, which (3) distributes work among server nodes. (4) After computation, results are accumulated by control and normalization. (5) Final results are sent back to client/pushed to cloud. (b) A schematic of the branching server framework. This method allows parallel accumulation of results between servers, reducing the number of data transfers from a linear relation to the number of servers, to a log relation.

several commercial products such as the Dropbox® and Microsoft® cloud storage system and is beyond the scope of this paper. During the treatment planning stage, a single server node is randomly assigned as a control server. The remote client initiates the MGCS process (step 2) by signaling the control server with required parameters. This control server is ultimately responsible for sending the dose results back to the client. The control server would intelligently distribute tasks to other computational servers in the MGCS network (step 3) based on the data size, the characteristics of the beam delivery (size, number, shape, orientation) and the number of convolution directions. For this purpose, a branching inter-server communication setup is employed, as shown in figure 3.1(b). The control server first communicates the dose computation parameters with the selected server nodes (in the figure 3.1(b) example setup, server nodes 1, 2, and 3). These servers then communicate with other servers in a similar manner. For the example setup, server nodes 2 and 3 communicate to server nodes 4 and 5 with the required dose computation parameters, and server node 5 then further propagates to server node 6. Once the computation is completed asynchronously by all the server nodes, data transfers of the 3D dose distribution occur from all servers back along their communication pipeline to the control server.

The data transfers occur in parallel in order to increase the network throughput. For the example setup in figure 3.2(b), the control retrieves the results from server node 1 (step 4), while simultaneously server node 2 retrieves and accumulates the results from server node 4, and server node 5 retrieves the results from server node 6. The control then moves on to the summed results of server node 2, while server node 3 retrieves the accumulated results from server 5. In step (5), the control server performs final 3D dose summation and normalization of results before sending them back to the client in the form of (a) DVH curves representing the dose to be delivered and (b) specific dosimetric endpoints for critical structures. Additionally, the 3D dose distribution computed in the cloud framework was also updated into the cloud-based data storage. This workflow is scalable for variations in the number of server nodes and GPUs per server.

Table 3.1. Definition of variables.

Number of server processes	N_S	Ratio of precomputed data size to results size	f_L
Number of devices per server process	N_G	Factor of acceleration from multiple streams	f_M
Result data size	d	Device memory allocation time	a_A
Compressed result data size	d_s	Device peer-to-peer set up time	a_S
Server read from disk speed	v_L	Device peer-to-peer retrieval time	a_R
Local server network speed	v_N	Data summation kernel time	k_S
Cloud/Client internet connection speed	v_I	Data normalization kernel time	k_N
Original single field computation time	T	Number of fields in dose calculation	n_F

Theoretical Model for the Multi-GPU Cloud Server Acceleration

A generalized methodology was developed to estimate performance gains for MGCS implementation with equations 2-15. The total computation, presented in equation 14, can be divided into six stages: (0) signal from client to control, (1) server nodes load precomputed data, (2) initialize GPU memory, (3) perform computations, (4) aggregate results, (5) consolidate results from server nodes to control, and (6) send final results back to client. Table 3.1 summarizes the variables in the below equations.

Let t_0 represent the time taken for the client signal to control and the subsequent chain of inter-server communication. Since the initial signaling between the client and the control does not include large data transfers, the value for t_0 was a few milliseconds. Let t_1 represent the time taken by the individual server nodes to load the 3D image dataset from cloud storage. It is defined as

$$t_1 = \frac{f_L d}{v_L}, \tag{2}$$

where d represents the total data size, f_L represents the ratio of the precomputed data size and the result data size, and v_L represents the disk read rate. Let t_2 represent the total time taken for initializing the device memory in each GPU (a_A), defined as

$$t_2 = a_A, \tag{3}$$

The computation time t_3 is defined as function of the time required for computing a single field dose distribution, T, multiplied by the total number of fields in the complete dose plan, n_F, divided by the product of the number of server nodes, N_S, and the number of GPUs per server node, N_G. It is defined as

$$t_3 = \left(\frac{n_F T}{N_S N_G}\right), \tag{4}$$

Let t_4 represent the time taken for the accumulation of the dose distribution results from the individual GPUs of each server process. It is defined as a product of the summation of peer-to-peer data retrieval time (a_R) and the GPU based dose summation time (k_S) and the number of additional GPUs utilized by the server process.

$$t_4 = (a_R + k_S)(N_G - 1) \tag{5}$$

Let t_5 represent the time taken for retrieving the results from all server nodes and normalizing the final result. It is defined as

$$t_5 = \frac{dN_s}{v_N} + k_S(N_S - 1) + k_N \tag{6}$$

Where v_N represents the inter-server network speed, k_S and k_N are the kernel execution times for summing and normalizing the results, respectively.

Finally, let t_6 represent the time taken to send the final computed results from the control server to the remote client, where v_I is the internet connection speed between the MGCS control server and the remote client.

$$t_6 = \frac{d}{v_I} \tag{7}$$

Algorithm Optimizations

Minimizing networks data transfers - inter-server communication. Unix sockets [24] were employed for inter-server communication. Establishing the socket connection was a simple procedure requiring only an auto-generated port number and the internet protocol address of the server node. The maximum data rate transfer on any network was fixed and so the only way to reduce the data transfer time was to minimize the number of data transfers, and the size of each data transfer. In order to minimize the number of data transfers, each server process was configured to run multiple convolutions and sum the results, which limited the memory transfers to one per server node. The workload distribution by the control was optimized to scale to any number of server nodes and equally distribute the convolution directions. To minimize the size of each data transfer only the sub-volume involved in the calculation was transferred as opposed to the full 3D data matrix. For a typical CT scan, over half of the image data is empty area outside of the patient's body, and only a small portion of the patient anatomy is actually involved in the dose computation. The anatomy involved in the calculation was defined by the pre-computed TERMA (total energy released in matter) matrix. The equations for terms t_2, t_4, t_5 and t_6 were modified to incorporate the compressed data size, d_c. Further reduction was achieved by converting the dose results from floating point (4 bytes) to short integer type (2 bytes). Extensive systematic studies were performed to test the accuracy of this method, and the results presented in §III.A confirm the expected precision of five significant figures. Terms t_5 and t_6 were reduced by a factor of two to reflect the change in data type.

Lastly, while increasing the number of server processes reduced the computation time, it also increased the time required to copy results back to the control. The bandwidth for control-node network data transfers saturated, which resulted in serialization of data retrieval. Therefore, instead of copying the data from each server node sequentially to the control, the branching method described from figure 1(b) was introduced. For three or more server nodes, a copy and summation may occur between two nodes concurrently with the control's data retrieval from a third node. This effectively reduced the number of copies to the control from N_s to $log(N_s)$. The logarithmic relationship was applied into term t_5.

Incorporating the above strategies into the terms t_2, t_4, t_5 and t_6 resulted in the following modifications:

$$t_2 = a_A\left(\frac{d_c}{d}\right), \tag{8}$$

$$t_4 = (a_R + k_S)(N_G - 1)\left(\frac{d_c}{d}\right) \tag{9}$$

$$t_5 = \frac{d_c \log N_S}{2v_N} + (k_S(N_S - 1) + k_N)\left(\frac{d_c}{d}\right) \tag{10}$$

$$t_6 = \frac{d_c}{2v_I} \tag{11}$$

Optimizing IPC methods - intra-server communication. The bandwidth for transferring data between the server processes was limited by the network interconnect speed. Copying memory using peer-to-peer access between GPUs has a much higher bandwidth than transferring memory between CPU processes. Accessing multiple GPUs through a single server process also requires less overhead than launching a server process for each device. Using this method, the convolution results were accessed peer-to-peer using CUDA IPC memory handles to consolidate the results from each GPU before copying back to the CPU in the parent server process and consolidating the results through inter-server communication. This provided the least amount of memory transfer on the CPU side, while maximizing the parallelism on the GPU side. The term t_2 was modified to include the CUDA IPC set up time, a_S, for each additional GPU as follows:

$$t_2 = \left(a_A + a_S(N_G - 1)\right)\left(\frac{d_c}{d}\right), \tag{12}$$

Optimizing memory concurrency – inter-GPU communication. Methods explored for optimizing memory concurrency included forked processes, multiple streams per GPU, multiple GPUs per process, and multiple machines. Only a single CUDA context may run on a GPU device at a time; launching multiple processes on a single device requires context switching. Context switching resulted in sequential kernel calls and additional overhead for scheduling tasks. Concurrent execution on the GPU was only possible using multiple streams within a single process. Operations in separate streams can overlap, while operations within a single stream will execute in order. Concurrent execution on the GPU using streams is limited only by the device resources, such as memory size and number of processing cores. However, some consideration must be given to avoid blocking calls, such as copying memory between the CPU and GPU, which will synchronize streams and diminish parallel performance. We thus modify the term t_3 as follows:

$$t_3 = f_M\left(\frac{n_F T}{N_S N_G}\right), \tag{13}$$

where f_M represents the factor of acceleration from the multiple stream usage.

Concurrency was further tested by forking the control process to communicate with each server node simultaneously, instead of looping over the server processes and signaling them sequentially. Similarly, the server process was forked for each GPU under its control, parallelizing the CPU-GPU communication and transferring data between devices using CUDA IPC memory handles. For this method, an array was allocated on GPU0 for each of the forked processes. Memory handles were created and copied into a mapped memory array accessible by all processes. After computation, the results were copied through peer-to-peer access.

Optimized model for total MGCS computation time. Applying the optimized representations for each term, the total time (t_{MGC}) of the MGCS implementation of an algorithm was defined as:

$$t_{MGC} = \frac{f_L d}{v_L} + \frac{f_M n_F T}{N_S N_G} + \left(\frac{d_c}{d}\right)[a_A + (a_S + a_R + k_S)(N_G - 1) + k_S(N_S - 1) + k_N] + \frac{d_c \log N_S}{2v_N} + \frac{d_c}{2v_I} \qquad (14)$$

The theoretical acceleration (A_{MGC}) using the MGCS framework was the ratio of the original algorithm's execution time and equation 16, giving the following final form.

$$A_{MGC} = \frac{n_F T}{t_{MGC}} \qquad (15)$$

When expanding to a multi-GPU implementation, a direct linear relationship between the acceleration and the total number of GPU devices being employed was expected. For the MGCS framework, the overhead of data transfers and other memory operations must also be considered. Therefore, the limit of acceleration should approach a direct linear relationship for computationally heavy tasks where the calculation time on a single GPU was much larger than the overhead of memory operations in the MGCS framework. This can be shown by rewriting equation 14 under the assumption that term 3 is much larger than the sum of all other terms, C:

$$t_{MGC} = \frac{n_F T}{N_S N_G} + C \approx \frac{n_F T}{N_S N_G};\ when\ \frac{n_F T}{N_S N_G} \gg C \qquad (16)$$

Substituting into equation 16 gives the theoretical limit for acceleration of the MGCS framework:

$$A_{MGC} = \frac{n_F T}{t_{MGC}} \approx n_F T \left(\frac{N_S N_G}{n_F T}\right) = N_S N_G \qquad (17)$$

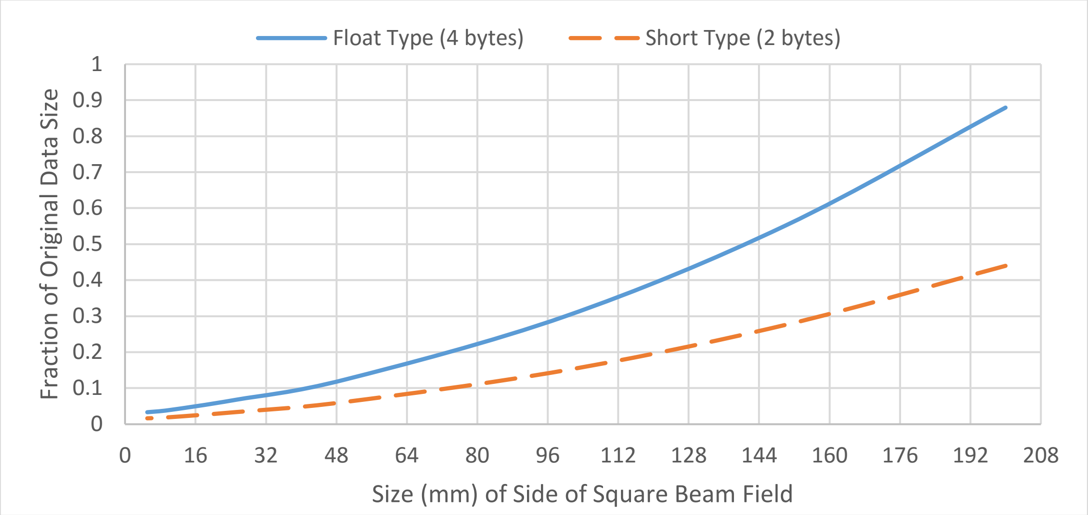


Figure 3.2. Potential reduction in data transfer size as a function of field size. Results show the fraction of the total volume for a 256 mm cubic data set of 1 mm isotropic voxels.

RESULTS

For numerical studies, the client process was run locally on 64-bit Linux with an 8-processor Intel Core i7 3.6GHz CPU. A set of three network server nodes were also running 64-bit Linux with four NVIDIA GeForce 980 GPUs running CUDA 6.5. An additional server node consisted of two NVIDIA GeForce 780 Ti GPUs, running CUDA 6.5. The server nodes through a 1 Gbps network interconnect.

Minimizing Networks Data Transfers - Inter-Server Communication

To reduce the data size and minimize data transfer time, the data was cropped around the active volume. The size of the cropped volume was determined by the pre-calculated TERMA volume plus an additional penumbra region to account for scatter. TERMA calculations generally completed in less than a millisecond. The original volume was a 256 mm sided cube, comprised of 1 mm isotropic voxels. The ratio of the reduced data size to the original data size followed a near linear fit as a

function of the cross-sectional area of the irradiation field. For example, for a 100x100 mm field, the data size could be reduced from 67 MB to 20 MB. The active volume included the main field as well as any voxels in the penumbra that would receive dose due to scattering. Converting from float to short data type further reduced the data size by a factor of 2, resulting in a reduction of

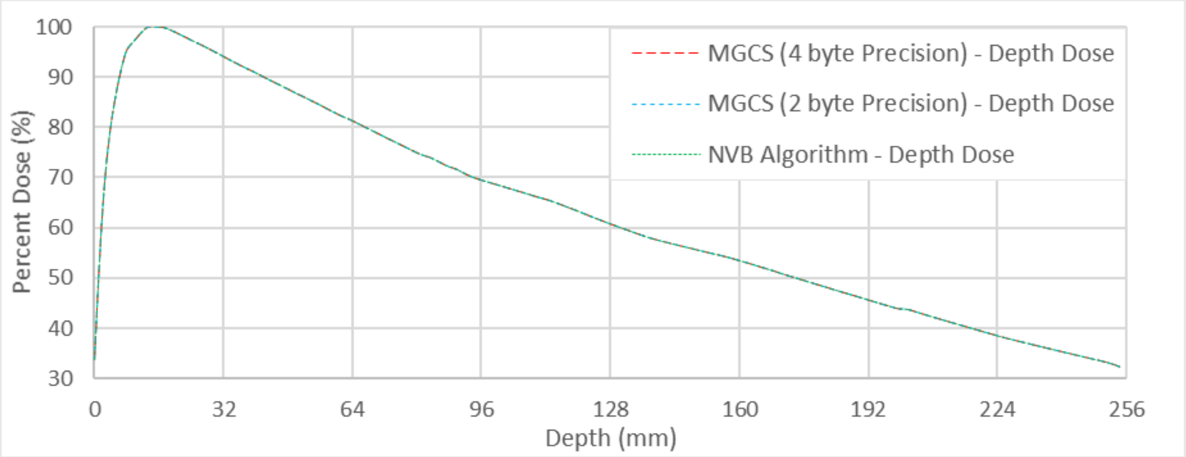


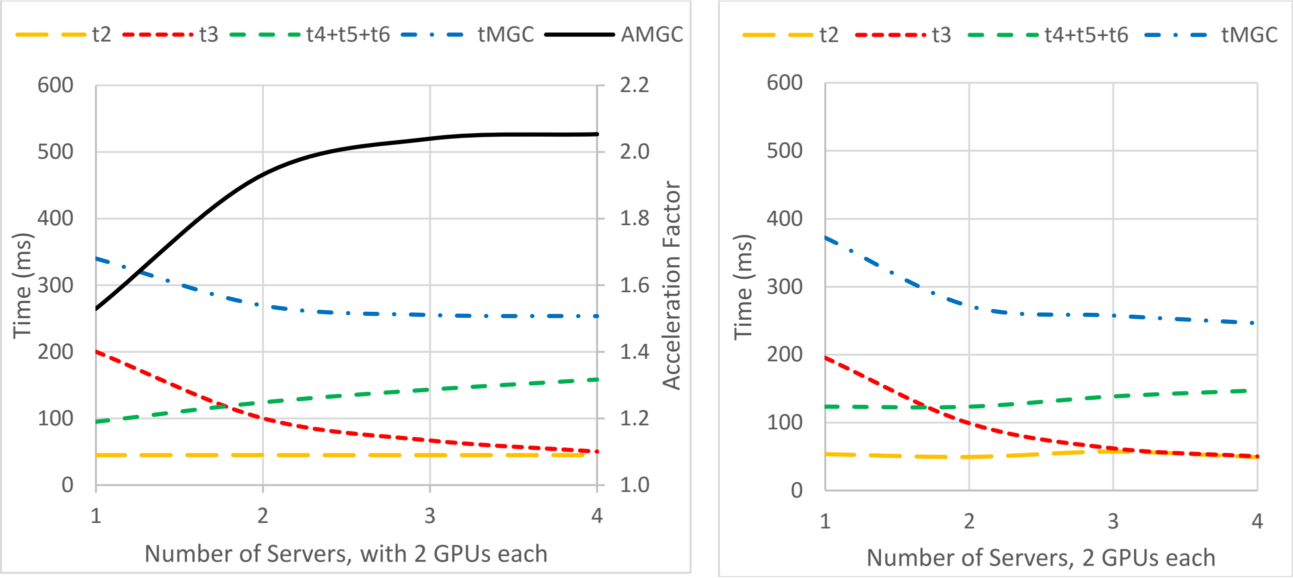
Figure 3.3. Accuracy of Multi-GPU Cloud Server Implementation. Depth dose curves for the original NVB algorithm, the 4-byte float precision MGCS implementation, and the 2-byte short precision MGCS implementation. Plots calculated using a 100x100 mm square field on a standard mediastinum phantom with 1 mm isotropic voxels and a size of 256 mm along each axis. The MGCS configuration depicted utilized 2 server nodes, with each node utilizing 2 GPUs.

85% for a 100x100 mm field. Figure 3.2 shows the potential reduction in data size for both float type and short type data.

Figure 3.3 shows a comparison of the depth dose profile through isocenter of a standard mediastinum phantom for the original NVB algorithm, the MGCS implementation using both floating point and short integer precision. There was minimal error when reducing precision, with a maximum of 4e-3%. The error between the NVB algorithm and the MGCS implementation along the depth dose curve was less than 1e-4%.

Numerical Analysis of the MGCS Acceleration

The observed results were also compared to the predicted results from the equations in section II.D. Table 3.2 shows the approximate values for GPU-specific (GTX 780 Ti) run time and network variables. The algorithm, specifically the convolution kernel, required too many GPU resources for concurrent execution, resulting in no performance gains from employing multiple streams, setting the acceleration factor, f_M, to 1.



(a) Model Predictions (b) Experimental Observations

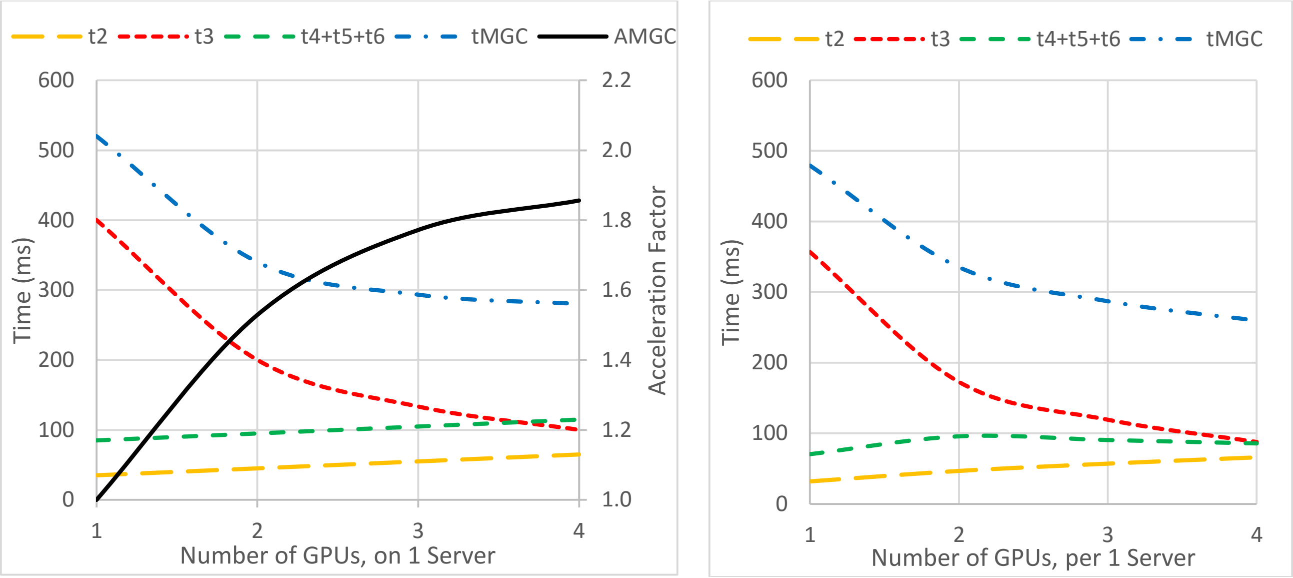
Figure 3.4. (a) Relationship between performance and increasing number of servers in the MGCS configuration for individual terms as predicted by the model. (b) Observed performance of individual terms when increasing the number of servers in the MGCS framework. Values reported for a single field dose calculation with parameters described in Table 2.

Table 3.3 shows the strong agreement between the predicted and observed results when increasing the number of GPUs on a single server process, with estimations within 5% of the observed performance. Using this simple calculation, many different combinations of resources can be compared to find the optimal distribution of the workload.

Table 3.2. Estimate values of variables for the NVB implementation on the multi-GPU Cloud Server Framework

f_M	1	a_A	35 ms
T	400 ms	a_S	10 ms
n_F	1	a_R	5 ms
v_N	0.125 MB/ms	k_S	5 ms
v_I	0.125 MB/ms	k_N	5 ms

Theoretically, increasing the number of server processes would reduce the total computation time by a linear proportion and this was seen in the convolution times (t3) of the model. Figure 3.4(a) shows the model predicted response for GPU initialization (t2), convolution computation (t3), and accumulation and normalization of the dose results (t4-t6) for a single field convolution calculation as a function of the number of servers in the MGCS framework, with each server employing two GPUs.



(a) Model Predictions (b) Experimental Observations

Figure 3.5. (a) Relationship between performance and increasing number of GPUs on a single server in the MGCS configuration for individual terms as predicted by the model. (b) Observed performance of individual terms when increasing the number of GPUs on a single server in the MGCS framework. Values reported for a single field dose calculation with parameters described in Table 2.

Figure 3.4(b) shows the actual timing data acquired from experiments. Similarly, figure 3.5 compares the predicted and observed response to an increasing number of GPUs being employed by a single server. The graphs show good correspondence between the predicted and observed values in both cases.

Table 3.3. Comparison between predicted and observed performance for a single server while increasing the number of GPUs.

Number of GPUs	Predicted Performance (ms)	Observed Performance (ms)	% Difference
1	485	479	1.24
2	330	335	1.52
3	288	287	0.35
4	274	260	5.11

Figures 3.4 and 3.5 also illustrate how memory operations can dominate for small computations, such as a single field dose convolution. Overall acceleration was inhibited by approximately equal contributions of the convolution calculations and memory operations to the total MGCS computation. Despite the additional latencies, the MGCS framework was still able to accelerate the single field dose convolution by 2x. This illustrates that the MGCS framework, with the optimization strategies

described above, was able to outperform a local single GPU system for even the most efficient, well-optimized processes. However, to fully utilize the increased computational power of the MGCS framework, the algorithm's computation time should be dominated by the calculation stage.

Optimizing Performance

For volumetric arc therapy (VMAT) or Tomotherapy planning, which regularly utilize 180-225 fields, the time required to calculate a complete dose distribution is much greater than the memory operations. Here, the MGCS framework approaches the linear acceleration limit, defined by the total number of GPUs. Figure 3.6 displays MGCS acceleration response to an increasing number of fields, when employing 1, 2, 3, or 4 servers, each utilizing 4 GPUs. For a 250 field dose plan, the MGCS framework reached peak accelerations of 3.98x, 7.88x, 11.7x, and 15.4x, for their respective number of servers. The biggest factors contributing to the maximum achievable acceleration were the original algorithm's computation time, and its proportion to the data transfer time as dictated by the size of the results. More computationally expensive algorithms such as many field dose calculations have the largest potential benefit from the MGCS framework, while minimizing the data transfers between the server nodes increased the framework efficiency.

DISCUSSION

Feasibility of MGCS Implementation

The feasibility of a MGCS framework for remote dose calculations was investigated in this paper. The proposed method could potentially enable clinics to make use of continually advancing technologies without the requirements of purchasing, maintaining, and frequently upgrading hardware. Additionally, a cloud-based solution holds the potential for much greater computing power than could feasibly be installed in the limited space of a clinic, and provide services to multiple clinics with the possibility for a measure of standardization across institutions.

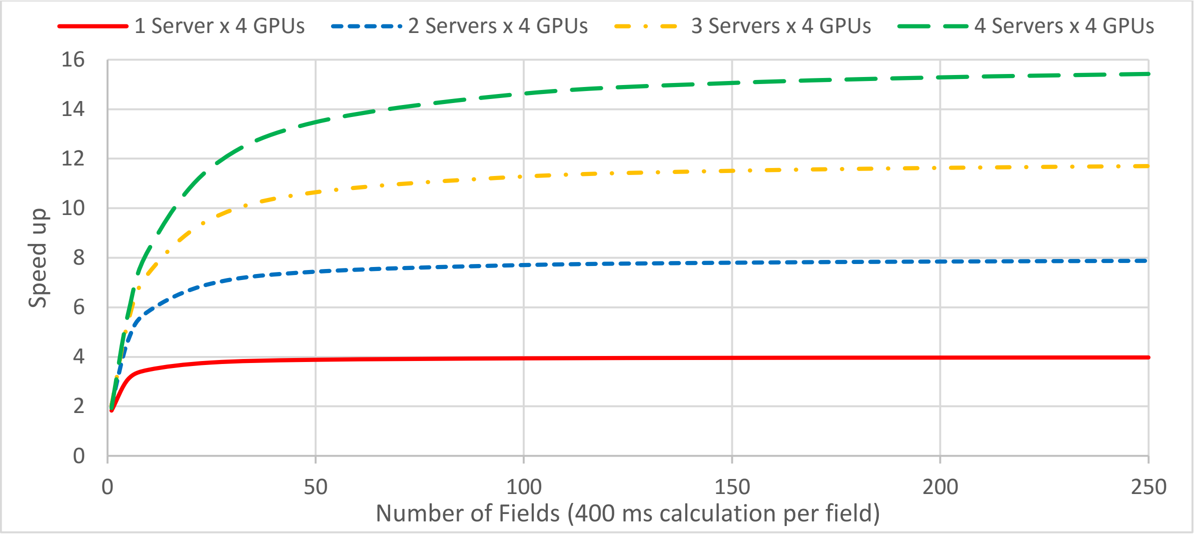


Figure 3.6. Acceleration as a function of the number of fields (each a 400 ms calculation on a single GPU) for different configurations of resources. Predicted maximum acceleration equal to the total number of GPUs.

A theoretical model to estimate the potential acceleration available using the MGCS framework was also presented. The model took into account the workload distribution in the MGCS framework, the original algorithm parameters, and the inter-GPU and inter-server communication latencies. The estimated computing time was numerically validated using a 14 GPU cloud computing setup. The model predicted performance and experimentally observed performance matched within 5%. In addition, discrepancies between the dose distributions calculated by a local, single-GPU implementation and the MGCS implementation were observed to be negligible.

For computationally heavy tasks, such as the many field dose plans of VMAT and Tomotherapy, the MGCS framework approaches the theoretical acceleration limit directly proportional to the total number of GPUs being utilized. For instance, for a 4 server configuration with 4 GPUs each (16 total GPUs), the acceleration of the MGCS framework peaked at 15.4x speed up. However, even for less intensive tasks, such as a single field dose convolution, the MGCS framework was still able to achieve 2x speed up, despite the additional memory operation overhead of instantiating a cloud-based solution.

The data transfer and other memory operation overhead prevent the MGCS framework from achieving the theoretical acceleration directly proportional to the total number of GPUs. One method investigated in this paper for mitigating these aspects was reducing the number and size of the data transfers. An alternative solution would be to improve the speed of inter-server communications. By replacing the 1 Gbps ethernet network with a 16 Gbps or PCIe3 interconnect backbone, data transfer times between the server nodes can be significantly reduced.

Clinical Implementation of MGCS and Future Work

Internet speed can be unreliable due to bandwidth limitations or slower connection speeds on the client end. In this case, DVH data and specific dose endpoints for critical targets and organs at risk could be sent directly back to the client, while the full dose distribution was synced with cloud storage. This would avoid any significant delays from transferring large amounts of data.

To maintain scalable performance for increasingly computationally heavy tasks, future work will focus on more efficiently distributing the workload among the server nodes. Distributing the workload of the incoming client task will require the control server to analyze the optimal number of required server processes and GPUs to utilize, using the model to estimate the best possible acceleration. The control will then query the server tree to identify available resources and idle server nodes. The control will also need to ensure the server nodes being queried have synchronized with the cloud, and already have access to all the necessary precomputed data. To do this, each task will be assigned a unique session identifier, which will have a flag associated with it to indicate the binary state of synchronization for this task. Additionally, data security and encryption will also be included to better quantify the feasibility of a cloud-based dose computing framework, and their subsequent effect on performance must be analyzed.

Applicability of MGCS Beyond Dose Computation

The feasibility of utilizing an MGCS framework for dose computations was investigated in this manuscript, but there are several other computationally expensive tasks in radiotherapy that could benefit from access to the power and throughput of an MGCS framework. Many treatment methodologies utilize an inverse optimization to determine beam arrangements. For conventional planning, the speed of such techniques is not a concern and can take several hours without interrupting the clinical workflow. However, for on-table adaptive re-planning, inverse optimization would need to complete in a matter of minutes. Distributing the work over several server nodes, and utilizing GPU parallelization could greatly accelerate the optimization process. More complex planning methods further emphasize the need for access to extensive computational networks to accelerate these tasks to a point where it becomes feasible to perform on-table adaptive re-planning. IMRT requires optimization of the motion of the multi-leaf collimators. Small physiological changes can completely alter the MLC sequence to better focus on the target anatomy. Another example is 4π treatment planning, which optimizes beam delivery in three dimensions, as opposed to the traditional axial in-plane optimizations. We envision a fully realized MGCS framework with meticulously parallelized algorithms will be able to perform daily adaptive adjustments, using the existing treatment plan as a priori information, within the time frame of a few minutes after the patient's daily imaging is acquired.

CONCLUSION

The MGCS framework that is presented in this paper may improve radiotherapy treatment outcomes by facilitating more frequent re-planning due to its potential for accelerating calculations. With the price-performance ratio getting smaller every day, more and more GPUs can be integrated with the cloud computing framework to progressively improve the computation speed. This would relieve the clinic of the obligations for purchasing, upgrading, and housing the hardware necessary for massively parallel computing. Any clinic could access the newest hardware and massive computational power of the cloud-based solution, providing significant accelerations for the most time intensive tasks that currently inhibit online adaptive therapies, allowing for a standardization of algorithms across institutions, and facilitating information sharing. The analytical model can be used as a preliminary test to determine which clinical algorithms will benefit the most from MGCS implementation and justify the development effort for conversion, in addition to guiding intelligent distribution of tasks within the MGCS framework. We believe the speed and accessibility advantages will make an MGCS framework integral to the future of radiotherapy, and specifically to the implementation of online ART into the daily clinical workflow.

REFERENCES

[1] Foroudi, F., Wong, J., Kron, T., Rolfo, A., Haworth, A., Roxby, P., Thomas, J., Herschtal, A., Pham, D., Williams, S., Tai, K. H., and Duchesne, G., "Online Adaptive Radiotherapy for Muscle-Invasive Bladder Cancer: Results of a Pilot Study," International Journal of Radiation Oncology Biology Physics 81(3), 765-771 (2011).

[2] Lindegaard, J., Fokdal, L., Nielsen, S., Juul-Christensen, J., and Tanderup, K., "MRI-guided Adaptive Radiotherapy in Locally Advanced Cervical Cancer from a Nordic Perspective," Acta Oncologica 52(7), 1510-1519 (2013).

[3] Nijkamp, J., Marijnen, C., Herk, M. V., Triest, B. V., and Sonke, J., "Adaptive Radiotherapy for Long Course Neo-adjuvant Treatment of Rectal Cancer," Radiotherapy and Oncology 103(3), 353-359 (2012).

[4] Schwartz, D., Garden, A., Thomas, J., Chen, Y., Zhang, Y., Lewin, J., Chambers, M., and Dong, L., "Adaptive Radiotherapy for Head-and-Neck Cancer: Initial Clinical Outcomes From a Prospective Trial," International Journal of Radiation Oncology Biology Physics 83(3), 986-993 (2012).

[5] Tuomikoski, L., Collan, J., Keyrilainen, J., Visapaa, H., Saarilahti, K., and Tenhunen, M., "Adaptive Radiotherapy in Muscle Invasive Urinary Bladder Cancer - An Effective Method to Reduce the Irradiated Bowel Volume," Radiotherapy and Oncology 99(1), 61-66 (2011).

[6] Capelle, L., Mackenzie, M., Field, C., Parliament, M., Ghosh, S., and Scrimger, R., "Adaptive Radiotherapy Using Helical Tomotherapy for Head-and-Neck Cancer in Definitive and Postoperative Settings: Initial Results," Clinical Oncology 24(3), 208-215 (2012).

[7] Qi, X. S., Santhanam, A., Neylon, J., Min, Y., Armstrong, T., Sheng, K., Staton, R., Pukala, J., Pham, A., Low, D., Lee, S., Steinberg, M., Manon, R., Chen, A., and Kupelian, P., "Near Real-Time Assessment of Anatomic and Dosimetric Variations for Head-and-Neck Radiation Therapy via Graphics Processing Unit-based Dose Deformation Framework," International Journal of Radiation Oncology Biology Physics 92(1), 415-422 (2015).

[8] Dong, P., Lee, P., Ruan, D., Long, T., Romeijn, E., Low, D., Kupelian, P., Abraham, J., Yang, Y., and Sheng, K., "4PI Noncoplanar Stereotactic Body Radiation Therapy for Centrally Located or Larger Lung Tumors," International Journal of Radiation Oncology Biology Physics 86(3), 407-413 (2013).

[9] Dong, P., Lee, P., Ruan, D., Long, T., Romeijn, E., Yang, Y., Low, D., Kupelian, P., and Sheng, K., "4PI Non-Coplanar Liver SBRT: A Novel Delivery Technique," International Journal of Radiation Oncology Biology Physics 85(5), 1360-1366 (2013).

[10] Neylon, J., Qi, X., Sheng, K., Staton, R., Pukala, J., Manon, R., Low, D. A., Kupelian, P., and Santhanam, A., "A GPU based high-resolution multilevel biomechanical head and neck model for validating deformable image registration," Med Phys 42(1), 232-43 (2015).

[11] Neylon, J., Sheng, K., Yu, V., Low, D., Kupelian, P., and Santhanam, A., "A Nonvoxel-based Dose Convolution/superposition Algorithm Optimized for Scalable GPU Architectures," Med Phys 41(10), 101711 (2014).

[12] Santhanam AP, Y, M., H, N., N, P., SL, M., and Kupelian, P., "A multi-GPU real-time dose simulation software framework for lung radiotherapy," International Journal of Computer Assisted Radiology and Surgery 7(5), 705-719 (2011).

[13] Constantin, M., Sawkey, D., Mansfield, S., and Svatos, M., "SU-E-E-05: The Compute Cloud, a Massive Computing Resource for Patient-Independent Monte Carlo Dose Calculations and Other Medical Physics Applications," Medical Physics 38(6) (2011).

[14] Miras, H., Jiminez, R., Miras, C., and Goma, C., "CloudMC: a cloud computing application for Monte Carlo simulation," Physics in Medicine and Biology 58, N125-N133 (2013).

[15] Poole, C., Cornelius, I., Trapp, J., and Langton, C., "Radiotherapy Monte Carlo simulation using cloud computing technology," Australasian Physical & Engineering Sciences in Medicine 35(4), 497-502 (2012).

[16] Poole, C., Cornelius, I., Trapp, J., and Langton, C., "Technical Note: Radiotherapy dose calculations using GEANT4 and the Amazon Elastic Compute Cloud," arXiv 1104(1408) (2011).

[17] Meng, B., Pratx, G., and Xing, L., "Ultrafast and scalable cone-beam CT reconstruction using MapReduce in a cloud computing environment," Med Phys 38(12), 6603-9 (2011).

[18] Kagadis, G. C., Kloukinas, C., Moore, K., Philbin, J., Papadimitroulas, P., Alexakos, C., Nagy, P. G., Visvikis, D., and Hendee, W. R., "Cloud computing in medical imaging," Med Phys 40(7), 070901 (2013).

[19] Moore, K. L., Kagadis, G. C., McNutt, T. R., Moiseenko, V., and Mutic, S., "Vision 20/20: Automation and advanced computing in clinical radiation oncology," Med Phys 41(1), 010901 (2014).

[20] D'Haese, P. F., Konrad, P. E., Pallavaram, S., Li, R., Prassad, P., Rodriguez, W., and Dawant, B. M., "CranialCloud: a cloud-based architecture to support trans-institutional collaborative efforts in neurodegenerative disorders," Int J Comput Assist Radiol Surg 10(6), 815-23 (2015).

[21] Silva, L. A., Costa, C., and Oliveira, J. L., "DICOM relay over the cloud," Int J Comput Assist Radiol Surg 8(3), 323-33 (2013).

[22] Silva, L. A., Costa, C., and Oliveira, J. L., "A PACS archive architecture supported on cloud services," Int J Comput Assist Radiol Surg 7(3), 349-58 (2012).

[23] Griebel, L., Prokosch, H., Kopcke, F., Toddenroth, D., Christoph, J., Leb, I., Engel, I., and Sedlmayr, M., "A scoping review of cloud computing in healthcare," BMC Medical Informatics and Decision Making 15(17) (2015).

[24] Stevens, W. R., [UNIX network programming], 2nd ed, Upper Saddle River, NJ: Prentice Hall PTR, v. <1-2 >, (1998).

CHAPTER 4: Near Real-time Assessment of Anatomic and Dosimetric Variations for Head-and-Neck Radiotherapy via a GPU-based Dose Deformation Framework

A version of this chapter has been published as a manuscript: Int J of Radiat Oncol Biol Phys, Vol. 92, No. 2, pp.415-422, 2015. doi: 10.1016/j.ijrobp.2015.01.033

ABSTRACT

Nearly real-time assessment of anatomic and dosimetric consequences for head and neck treatment is feasible using a graphics processing unit-based deformable registration framework. Substantial interfraction anatomic changes resulting in clinically relevant dosimetric variations were observed for 11 head and neck cases. Although the cumulative target mean and maximum doses varied insignificantly, the cumulative minimum target and parotid gland doses deviated significantly from the plans. Clinical implementation of this technology may enable timely plan adaptation and potentially lead to improved outcome.

Purpose. The purpose of this study was to systematically monitor anatomic variations and their dosimetric consequences during intensity-modulated radiotherapy (IMRT) for head-and-neck (H&N) cancer using a graphics processing unit (GPU)-based deformable image registration (DIR) framework.

Methods. Eleven IMRT H&N patients undergoing IMRT with daily megavoltage CT (MVCT) and weekly kilovoltage CT (kVCT) scans were included in this analysis. The pre-treatment kVCTs were automatically registered with their corresponding planning CT through a GPU-based DIR framework. The deformation of each contoured structure in the H&N region was computed to account for non-rigid change in the patient setup. The Jacobian determinant of the PTVs and the surrounding critical structures were used to quantify anatomical volume changes. The actual delivered dose was calculated accounting for the organ deformation. The dose

distribution uncertainties due to registration errors were estimated using a landmark based gamma evaluation.

Results. Dramatic interfractional anatomic changes were observed. During the treatment course of 6-7 weeks, the parotid gland volumes changed up to 34.7%, and the center-of-mass displacement of the two parotid glands varied in the range of 0.9-8.8 mm. For the primary treatment volume, the cumulative minimum/mean/EUD doses assessed by the weekly kVCTs were lower than the planned doses by up to 14.9%(p=0.14), 2%(p=0.39), and 7.3%(p=0.05), respectively. The cumulative mean doses were significantly higher than the planned dose for the left-parotid (p=0.03) and right-parotid glands (p=0.006). The computation including DIR and dose accumulation was ultra-fast (~ 45seconds) with registration accuracy at the sub-voxel level.

Conclusions. A systematic analysis of anatomic variations in the H&N region and their dosimetric consequences is critical in improving treatment efficacy. Near real-time assessment of anatomic and dosimetric variations is feasible using the GPU-based DIR framework. Clinical implementation of this technology may enable timely plan adaption and improved outcome.

INTRODUCTION

Intensity modulated radiation therapy (IMRT) is a standard treatment technique for head-and-neck (H&N) cancer. IMRT has demonstrated the capability of delivering highly conformal doses to targets while sparing adjacent critical structures including parotid glands, spinal cord, etc. Daily volumetric image guidance not only improves patient alignment and dose delivery accuracy, but also reveals patient anatomic changes resulting from patient weight loss, tumor shrinkage, soft tissue deformation, and internal organ motion [1, 2]. These anatomic changes are commonly observed among H&N patients undergoing radiotherapy [1-5]. If unaccounted for, they may have detrimental effects on tumor control and/or organ-at-risk (OAR) sparing [1, 3, 4].

Adaptive radiotherapy (ART) is an appealing concept that aims to adjust the treatment plan based on the anatomical changes assessed on a daily basis using pre-treatment volumetric images [6-12]. Wu et al. [3] reported that the dosimetric benefit of replanning with reduced margins could result in up to 30% parotid gland dose sparing. Lee et al. [4] reported an average of 15% parotid mean dose difference between the delivered versus the planned doses due to anatomic changes during a course of radiation treatment. Recently, Schwartz et al. [9] performed a prospective adaptive trial for a group of 24 H&N cancer patients with 1-2 replan(s) in the middle of the treatment course. The early outcomes indicated promising clinical outcome results including low initial toxicity and high disease control. Chen et al. [13] also concluded that ART conveys a significant benefit in appropriately selected patients with H&N cancer. However, clinical implementation of ART remains challenging and labor intensive due to the complexity and lack of robustness in automated image registration/segmentation/dose summation. Subsequently, the integration of ART in H&N treatment is mostly manual and empirical without precise knowledge of the most appropriate timing and frequencies to initiate ART [1-4]. A robust automated ART framework is essential to implement the concept in routine clinical workflow without inducing treatment delay or excessive staff burden.

We aimed to validate an in-house deformable image registration (DIR) and dose accumulation framework [14] that registers the patient's daily treatment scan to the planning CT using a patient-specific biomechanical head & neck model and a multi-resolution registration method with the following goals: 1) enable fast assessment of the anatomic changes and organ motion for both targets and OARs during the course of treatment; 2) evaluate the resulting dosimetric differences between the delivered and the planned doses. Our ultimate goal is to monitor the delivered dose to the primary targets and critical structures in nearly real-time and to facilitate a data-drive decision-making process for ART.

MATERIALS and METHODS

In-House GPU-Based Dose Deformation and Accumulation Framework

The General Purpose Graphics Processing Units (GPU) based framework mainly involved a multi-resolution optical flow registration algorithm for registering simulation CT with corresponding weekly CT datasets [14, 15]. The computational steps were optimized to ensure the registration algorithm is completed in sub-minute computational time. First, the target volumes and OARs delineated on the planning CT were registered in a non-rigid manner and transferred to the weekly CT images. The deformation of each contoured structure was computed to account for non-rigid changes in the patient setup. Secondly, warping the planning kVCT anatomy to the weekly anatomy, a new warped kVCT was generated. The planning dose distribution was then overlaid with the warped kVCT. To compute the dose delivered to each voxel in the planning volume, the deformation map was used to accumulate the overlaid dose back on the planning CT. This generated a new dose map that corresponded to the underlying anatomy in the weekly CT. The new contour was created automatically by taking each voxel in the new data volume, and mapping it back to the planning CT using the deformation vector. The deformed contours allow for dose calculation and accumulation, resulting in dose–volume histograms (DVHs) and other dosimetric parameters. Thirdly, the Jacobian determinant for the PTVs and the critical structures were used to quantify anatomical volume changes for each week. Fourthly, a gamma analysis [7] was performed to provide a quantitative comparison between the calculated doses with respect to the planning dose distribution. The acceptance criteria for the gamma test were set to 1% intensity difference in 1 mm^3 neighborhood range, gamma≤1 was considered acceptable.

Landmark Based Registration Validation

The key of the work was the accuracy and the robustness of the in-house registration algorithm. The validation of the proposed DIR framework, including the deformable image registration and the dose integration were performed using planning kV and weekly kV images. A landmark-based interactive validation tool was developed to evaluate the uncertainty in dose distribution due to registration error. Consider the planning kVCT to be the source (or reference) image and the final week of the weekly kVCT to be the target image. For each of the selected landmarks in the reference image, the corresponding landmark in the target kVCT data was calculated using the image registration algorithm and visually displayed as cross hairs in the target image. A set of 100 landmarks were selected on the target/critical structures of a reference kVCT and mapped to the target image. Once the landmarks were picked, the user either accepted the registration results or marked the correct landmark on the target image. The Target Registration Error (TRE) metrics [14, 15], defined as the sum of squared difference between the ground truth displacement and the displacement computed from the registration process, for each of the datasets were computed.

Clinical Data

Eleven H&N patients treated with simultaneous integrated boost (SIB) technique on a Hi-ART Tomotherapy system (Accuray Inc., CA) were included to validate the in-house developed framework. Patient data for this study were acquired as part of an IRB-approved adaptive planning protocol. All patients received two sets of volumetric image scans during treatment: weekly kVCTs and daily MVCTs acquired before each treatment. The planning kVCT image set was acquired prior to the start of treatment on a Philips Brilliance CT system (Philips Medical Systems, Best, The Netherlands). Weekly volumetric kVCT images of each patient enrolled in the protocol were re-acquired on the same equipment throughout the course of treatment. All patient kVCT images were acquired with the patient in the simulated treatment position using a 50-70 cm FOV, 512×512 in-plane resolution, and 3 mm slice thickness. Table 4.1 shows the patient characteristics. A total of 71 weekly kVCT scans were acquired and analyzed. Patients 9, 10 and 11 (patient numbers are shown in Tables 1 and 2) were re-planned during the middle of the treatment. MVCTs were not used in this study due to poorer soft tissue contrast. Unless otherwise specified, the accumulative doses were computed based on weekly kVCTs.

The prescription doses were 2.0-2.1 Gy/fx in 30-35 fractions. The targets and critical structures, such as CTVs, PTVs, spinal cord, and parotid glands were delineated by a radiation oncologist on the planning CT. A 3 mm margin was used for the CTV-to-PTV expansion, and a 5 mm margin was applied to the cord to account for setup uncertainty. All patients were treated with TomoTherapy helical IMRT (version V4.1), with a field size of 2.5 cm, pitches of 0.277-0.3 and modulation factors of 2.2-3.2.

Table 4.1. Patient characteristics and treatment delivery summary

Pt	Diagnosis	No. of fx	Date enrolled (mo/d/y)	Tx beginning (mo/d/y)	Tx completion (mo/d/y)	Tx duration (d)	Initial weight (kg)	Final weight (kg)	Weight change (kg)
1	Nasopharynx	33	06/23/11	06/23/11	08/09/11	47	81.6	76.2	-5.4
2	Tonsil	35	11/17/11	11/17/11	12/30/11	43	93.4	83.8	-9.6
3	Tonsil	35	11/25/09	12/07/09	01/26/10	50	88.0	85.9	-2.1
4	Tonsil	35	04/23/09	05/07/09	07/02/09	56	95.7	84.8	-10.9
5	Tonsil	35	10/20/10	10/28/10	12/20/10	53	72.1	63.3	-8.8
6	BOT*	35	12/29/10	01/07/11	02/18/11	51	63.5	62.6	-0.9
7	Tonsil	35	08/17/09	08/31/09	10/09/09	39	83.9	76.9	-7.0
8	Tonsil	30	10/08/12	10/16/12	11/28/12	43	86.4	82.0	-4.4
9	Tonsil	35	11/21/11	11/21/11	01/05/12	45	107.5	96.0	-11.5
10	BOT*	35	10/18/12	10/22/12	12/11/12	50	99.8	89.1	-10.7
11	Ethmoid Sinus	35	03/02/11	03/02/11	04/20/11	49	84.8	79.4	-5.4

Abbreviations: Pt = patient; BOT = base of tongue; fx = fraction; Tx = treatment

Assessment of Anatomic and Dosimetric Variations

The anatomic and positional changes for the targets and the parotid glands were measured on the weekly kVCT scans. For the targets, the delivered mean/minimum/maximum dose, D90, D95, V90, V95 and V100, as well as the equivalent uniform dose (EUD) were calculated (assuming □=-15 for the targets) and collected. For the OARs, we considered the minimum/mean/maximum dose for the cord and parotid glands. The center-of-mass (COM)

for the PTV1, parotid glands, and the COM distances between these structures were measured. Weekly delivered doses were estimated assuming constant anatomy for that week as reflected by the weekly kVCT. Finally, the accumulated dose was calculated and compared to the planned dose.

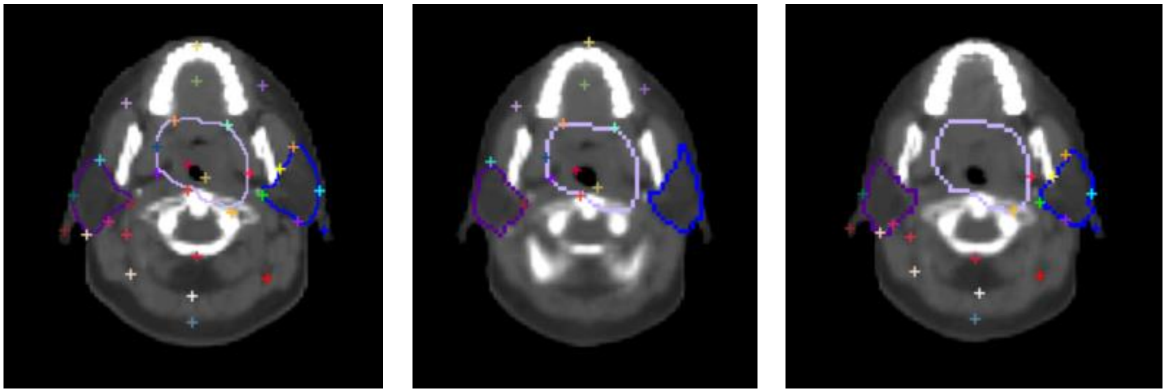

(a) (b) (c)

Figure 4.1. Verification of in-house deformable image registration, using a landmark tool. (a) Source image with the delineated structures; (b and c) target image overlaid with the deformed contours. The selected landmark points in the source and target images were displayed as crosshairs.

RESULTS

Accuracy and Robustness of the Framework

Figure 4.1 shows the verification of the in-house DIR framework using a landmark tool. (a) Source image (the planning CT) with delineated structure outlines of the target, left- and right- parotids. (b) and (c) Target image (a weekly kVCT) overlaid with the deformed contours. The corresponding landmark points in the target image were calculated using the image registration algorithm and visually displayed as cross hairs.

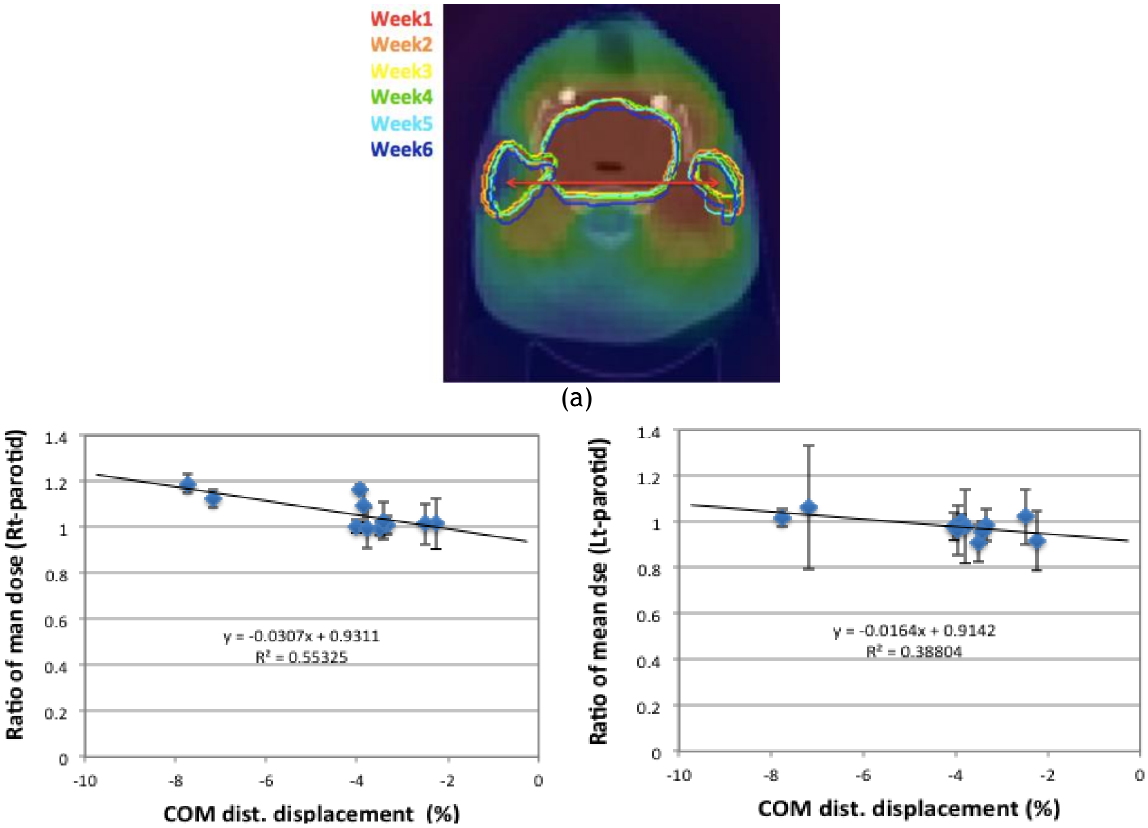
Table 4.2. Average ±SD registration accuracy across the entire body.

	Overall Accuracy (mm)		PTV Accuracy (mm)		Left Parotid Accuracy (mm)		Right Parotid Accuracy (mm)	
Subject	Avg	±SD	Avg.	±SD	Avg	±SD	Avg	±SD
1	1.00	0.71	0.68	0.45	1.87	0.97	1.63	0.77
2	1.08	0.78	1.05	0.42	2.04	1.09	1.70	0.85
3	1.23	0.84	1.69	0.70	1.97	0.81	2.06	0.87
4	1.01	0.54	1.21	0.50	1.32	0.51	1.45	0.54
5	0.99	0.62	1.23	0.55	1.28	0.50	1.51	0.68
6	0.98	0.73	1.52	0.74	1.69	0.82	1.73	0.79
7	0.88	0.64	0.74	0.39	1.95	1.13	1.87	1.03
8	0.90	0.82	2.42	1.53	1.97	1.20	1.90	1.16
9	0.99	0.76	1.88	1.10	1.68	0.93	1.58	0.68
10	1.02	0.72	0.97	0.41	2.13	1.17	2.11	1.16
11	1.25	0.79	1.34	0.56	1.84	0.70	1.95	0.80

Table 4.2 shows the registration accuracy and the standard deviation of the whole body, PTV and left/right-parotids using a landmark-based TRE metric (14). The averaged TREs were in the range of 0.88–1.25 (average 1.03±0.72) mm, 0.68–2.42 (average 1.34±0.67) mm, 1.28–2.13 (average 1.79±0.89) mm and 1.45–2.11 mm (average 1.77±0.85) mm for entire patient anatomy, PTV and left- and right parotids, respectively. Given the pixel size of 1.95×1.95×3 mm, the proposed DIR framework reached sub-voxel accuracy.

We further evaluated the dose distribution uncertainties due to registration errors. The landmark based interactive tool was developed to evaluate the uncertainty of registration error. We modified the gamma dose distribution evaluation tool to quantify the effect of the spatial uncertainty of the deformable registration on dose distribution. Using the mean and standard deviation of the target registration error, we introduced a normally distributed random displacement during evaluation. For each voxel in the test dose distribution, the Gamma analysis works by finding the corresponding voxel in the truth dose distribution, and performing a local neighborhood search to evaluate the euclidean magnitude in dose/distance space. The random displacement was applied when finding the corresponding voxel in the truth dose distribution, to effectively shift the local search neighborhood. The Gamma analysis was then performed as usual. The introduction of random error caused less than a 1% increase in the percentage of voxels that failed both the gamma evaluation (1%/1mm), and a direct dose comparison.

Multiple observers, including the primary physician, physician residents and physicists reviewed the deformed contours for the target and parotid glands using the proposed landmark verification tool. Minimal variations (1-2 mm) of inter-observers errors were found, which is comparable to the TRE metrics.



(a)

(b) (c)

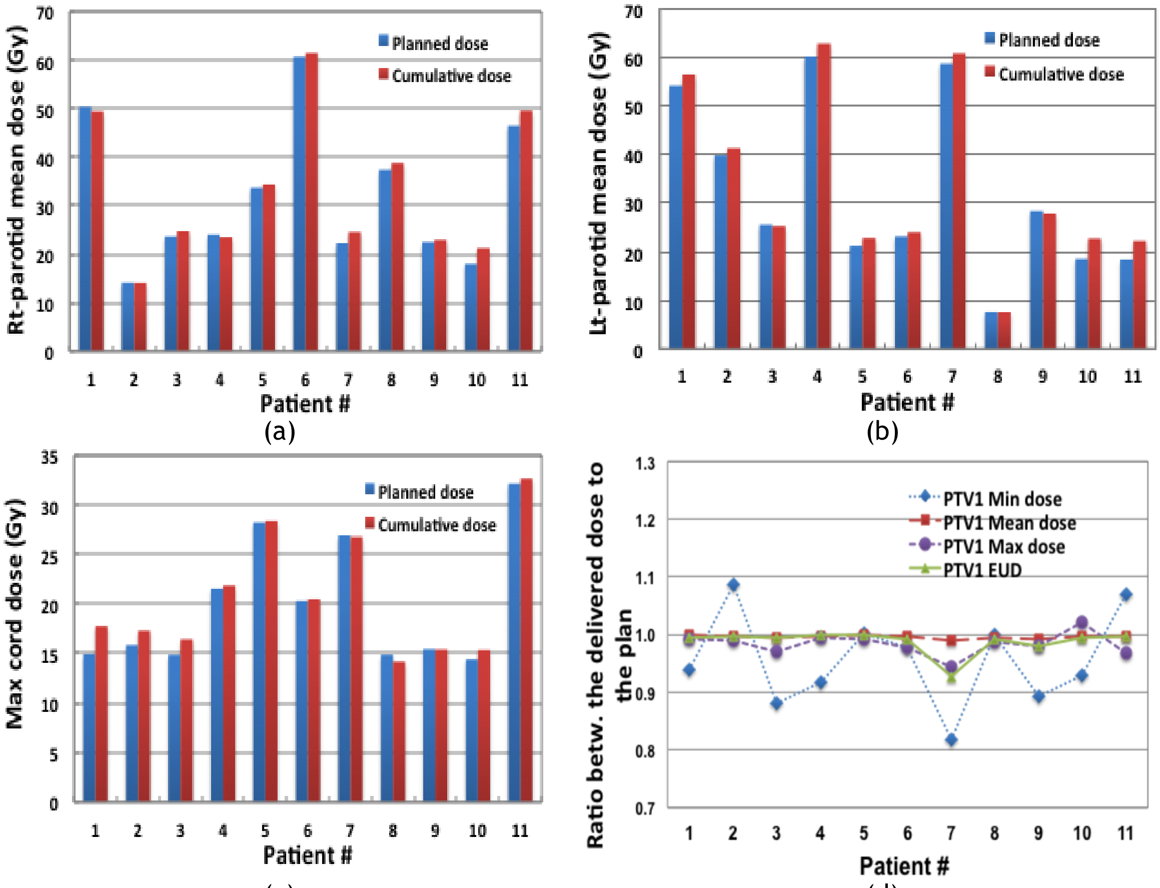
Figure 4.2. (a) Contours of the targets and parotid glands assessed from the weekly kVCT scans overland on the planning dose distribution. Correlations between the displacement of the COM distances of the parotid glands and the mean dose variation at the end of the treatment for the right (b [Rt]) and left (c [Lt]) parotid.

Abbreviations: COM = center of mass; dist. = distance; kVCT = kilvoltage computed tomography.

Interfractional Variations

The in-house tool was applied to analyze the patient cohort of 11 cases. The volume changes were assessed by weekly kVCT scans and normalized to the planning volume. The volume changes varied from -34.7 to 14.6% and -27.7 to 12.6%, respectively, for the left and right parotids during the 6-7 week treatment course. The volume increases between the planning (wk0) CT to the week 1 (wk1) CT were likely due to the elapsed days between the planning CT and the start date of the 1st treatment (Table 4.1).

Figure 4.2 (a) shows the contours of the parotid glands throughout the treatment course of 6 weeks for a representative patient. On average, the COM distances between the two parotid



(a) (b)

(c) (d)

Figure 4.3. Comparison of cumulative doses to planned doses for (a) right (Rt) parotid (*P*=.03), (b) left (Lt) parotid (*P*=.006), and (c) maximum cord doses compared to planned doses (*P*=.18). (d) Ratios of cumulative doses normalized to planned dose are shown the PTV1.

Abbreviations: EUD = equivalent uniform dose; PTV = planning target volume.

glands appeared to reduce in the range of 0.9 to 8.8 mm (mean: 4.9±2.3) mm at the end of the treatment course, meaning that the parotid glands were gradually moving toward the patient mid-plane. Figure 4.2 (b) and (c) shows the COM distance (normalized to the COM distance at the planning stage) versus the ratio of the mean doses normalized to the planned mean doses for the parotid glands. The delivered mean doses increased as the parotid glands gradually migrated towards the mid-sagittal (high-dose region) plane. Linear regression was performed for the mean parotid gland doses as a function of COM displacement. Moderate correlation was observed between COM displacement and the mean parotid dose deviation from the plan. This observation was consistent with the published literature [2, 4, 11, 12].

All patients lost weight over the treatment course (Table 4.1). The average and relative weight losses were 7.0±3.6 kg and 7.8±3.7%, respectively. Given the treatment duration of 6-7 weeks, the weight loss per treatment elapsed days was approximately 0.14±0.09 kg/day. Linear regression (R^2<0.4) shows there was mild correlation for patient weight loss with mean parotid dose change, but no apparent correlation was found between the cord maximum dose, the PTV1 mean dose and patient weight change (not shown).

The cumulative mean doses assessed by the weekly kVCT scans for the PTV1 were 68.9±6.1 Gy versus 68.8±6.2 Gy for the planned dose (p=0.39 using paired t-tests). The maximum cord dose delivered was 43.7±7.5 Gy compared to 40.7±4.2 Gy (p=0.18). However, significantly higher mean doses were seen for both parotid glands in the composite plans (p=0.03 for the left parotid and p=0.006 for the right parotid) shown in Figure 4.3 (a-b).

Figure 4.3(d) shows the ratio of the cumulative dose to the planned dose for the PTV1 in this patient cohort. While the maximum doses were consistent with the planned maximum doses within 5.7%, the cumulative mean dose ratios were within 1.1% of the planned mean doses for PTV1. Target DVHs also showed a moderate level of variation. Cold spots were observed in the cumulative dose distributions for the PTV1 in 6 out of 11 patients. Up to 14.9% of minimum dose reduction was observed for Patient 7 (who had the second largest PTV volume), resulting in significant EUD changes (p=0.05) from the plan.

Figure 4.4 shows (a) the plan and the deformed structures on a weekly kVCT scan for a representative case; (b) the planned dose distributions; (c) the delivered dose distribution; (d) the calculated gamma distribution for those voxels with gamma>1 on the weekly kVCT scan. Such gamma maps can be used to identify areas that need closer inspection. At looser gamma criteria of 2 mm/2%, 71.7% (right-parotid) and 89.7% (left-parotid) volume saw changes in dose, but such dose changes weren't extreme since the failure rates were minimal at 3%/3mm (failure rates of 0% for the right and 1.4% for the left-parotid). (e) Comparison of

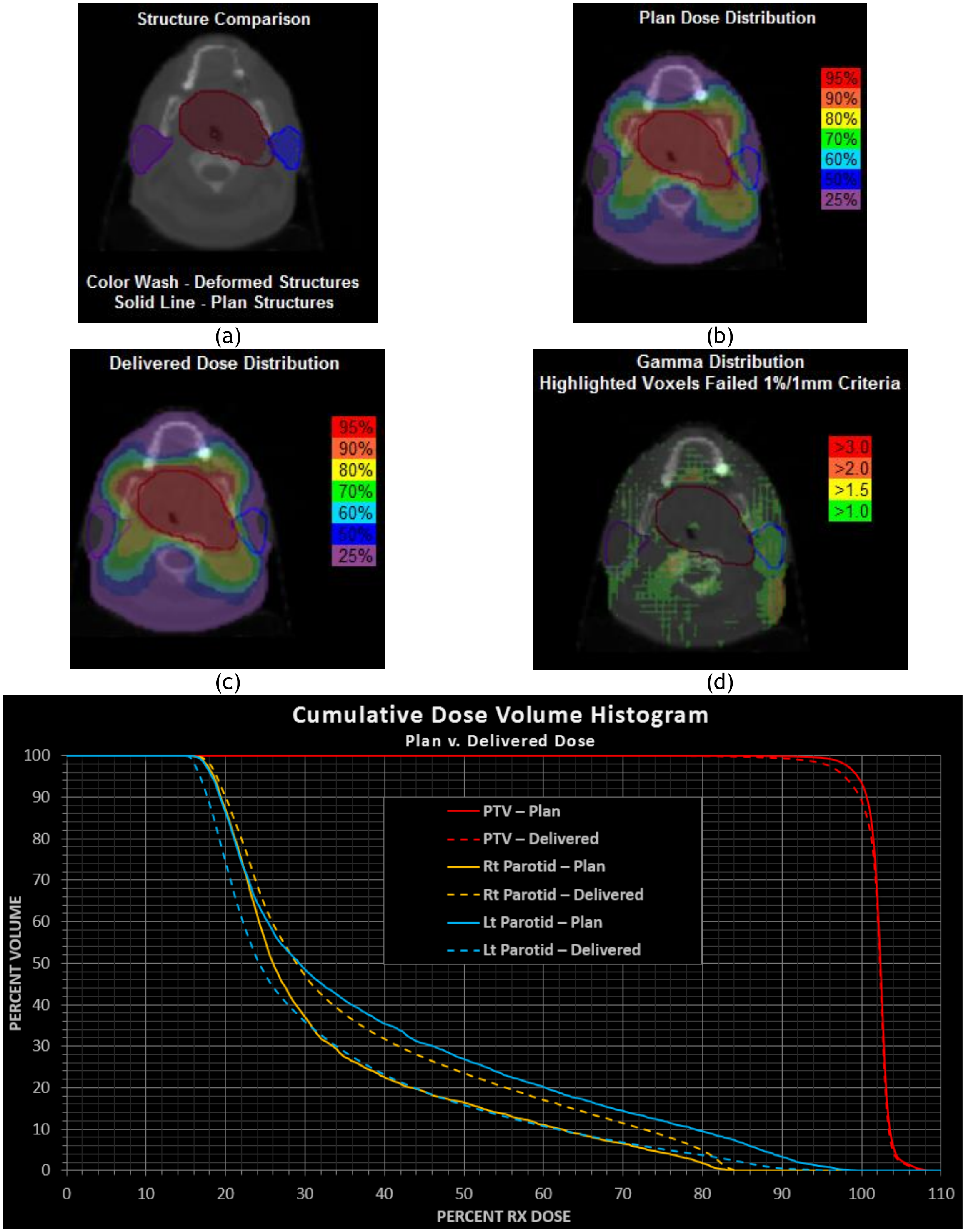


Figure 4.4. (a) Planned and deformed structures on a weekly kVCT; (b) planned dose distributions; (c) delivered dose distribution; (d) calculated gamma distribution (gamma > 1) overlaid on the weekly kVCT; (e) comparison of the planned to delivered DVHs for the PTV1 and parotid glands.

Abbreviations: DVH = dose-volume histogram; kVCT = kilovoltage computed tomography; Lt = left; PTV = planning target volume; Rt = right.

the plan and delivered DVHs for the PTV1 and parotid glands. Most of the gamma failure is around the surface of the patient potentially due to weight loss and minor posture changes.

Run-Time Analysis

The in-house dose deformation & accumulation tool achieved a fast calculation of 45 seconds for registering one weekly kVCT with a planning CT, including 1) data resizing and resampling of approximately 2 seconds (resampled data dimensions: 200×200×50 voxels, resampled voxel dimensions: 1.95×1.95×3.0 mm); 2) The deformable image registration using optical flow registration algorithm of 20 seconds; 3) Jacobian analysis of 6 seconds; 4) and gamma analysis of 5 seconds and 5) other minor processes such as file reading and writing of 12 seconds [14].

DISCUSSION

A near real-time anatomic and dosimetric assessment and evaluation framework was presented that facilitates clinical decision-making for ART by quantitatively accounting for plan quality degradation during the treatment course. A quantitative patient-specific biomechanical H&N anatomic model assembled using the conventional CT simulation (to account for subject specific sub-anatomy locations), was employed to register with routine on-board CT (to monitor the effects of posture/physiologic variations in gross treatment volume).

Progressive anatomical changes during the treatment resulted in substantially increased doses to the parotids and/or potential cold spots to the targets [1-4]. The dosimetric degradation was a result of the compounding factors including the percent of volume and positional changes for the parotid glands, tumor shrinkage and patient weight loss, etc. Larger weight loss may result in the larger COM reduction of the parotid glands, which led to larger delivered doses to the parotids, but the correlation was not strong. Overall, our results are consistent with previous studies demonstrating dramatic patient anatomic changes during the radiation treatment for H&N cancer [9, 10, 16].

The concept of performing ART on a regular basis to quickly compensate target underdoses and/or normal structure overdoses is appealing. However, its implementation is challenging due to the prohibitively labor intensive and time consuming required to delineate and validate the target and OARs on a daily basis. ART usually involves altering the planned doses according to variations in patient anatomy. This relies on an accurate representation of the changing dose distribution within the patient, which generally requires a full dose recalculation. To further reduce online re-planning time, this work adopted a dose resampling/warping method (of the planned dose distribution) to assess three-dimensional dose distribution at the time of treatment delivery. The dosimetric differences were validated against full dose recalculation for prostate and H&N radiotherapy by previous publications, and proven to be an acceptable (within ±5%) and effective implementation in current clinical practice [17-20]. Furthermore, although the image quality for various on-board imaging modalities (such as kilo-voltage and mega-voltage CBCTs and megavoltage CT) was sufficient for bony landmark based patient alignment purposes [2, 5, 21], they generally yield inferior image quality that could reduce the image registration, segmentation and dose deformation accuracy for adaptive planning. These limitations further underline the importance of an automated framework that is inter-observer dependent and robust to different imaging qualities.

Next step is to integrate the framework into our clinical workflow to 1) monitor the actual dose delivered to the primary targets and critical structures in a systematical manner and 2) to flag large dose degradation between the planned and delivered dose distribution to trigger a detailed plan reviewing process and/or a potential plan adaption. Given a large number of

H&N patients are being detected yearly [22] versus the limited clinical resources, it is not realistic (and probably not necessary) to apply ART to all H&N patients. The presented tool may efficiently identify a subset of H&N patients for whom ART are most beneficial. In the long run, a longitudinal study for a randomized patient population may shed light to establish the standardized adaptive protocol.

CONCLUSION

We demonstrated the feasibility of an ultra-fast assessment and documentation framework for systemically monitoring the anatomic and dosimetric variations during the course of H&N treatment. The delivered mean dose to the target appeared to be consistent with the plan, while the cumulative EUDs for the PTV1, cumulative dose to the OARs may be significantly different from the planned doses. The automated framework may offer timely interventions such as ART. Clinical implementation of this technology may lead to improved outcome.

REFERENCES

[1] Barker, J. L., Jr., Garden, A. S., Ang, K. K., O'Daniel, J. C., Wang, H., Court, L. E., Morrison, W. H., Rosenthal, D. I., Chao, K. S., Tucker, S. L., Mohan, R., and Dong, L., "Quantification of volumetric and geometric changes occurring during fractionated radiotherapy for head-and-neck cancer using an integrated CT/linear accelerator system," Int J Radiat Oncol Biol Phys 59(4), 960-70 (2004).

[2] Gregoire, V., Jeraj, R., Lee, J. A., and O'Sullivan, B., "Radiotherapy for head and neck tumours in 2012 and beyond: conformal, tailored, and adaptive?," Lancet Oncol 13(7), e292-300 (2012).

[3] Wu, Q., Chi, Y., Chen, P. Y., Krauss, D. J., Yan, D., and Martinez, A., "Adaptive replanning strategies accounting for shrinkage in head and neck IMRT," Int J Radiat Oncol Biol Phys 75(3), 924-32 (2009).

[4] Lee, C., Langen, K. M., Lu, W., Haimerl, J., Schnarr, E., Ruchala, K. J., Olivera, G. H., Meeks, S. L., Kupelian, P. A., Shellenberger, T. D., and Manon, R. R., "Assessment of parotid gland dose changes during head and neck cancer radiotherapy using daily megavoltage computed tomography and deformable image registration," Int J Radiat Oncol Biol Phys 71(5), 1563-71 (2008).

[5] Qi, X. S., Hu, A. Y., Lee, S. P., Lee, P., DeMarco, J., Li, X. A., Steinberg, M. L., Kupelian, P., and Low, D., "Assessment of interfraction patient setup for head-and-neck cancer intensity modulated radiation therapy using multiple computed tomography-based image guidance," Int J Radiat Oncol Biol Phys 86(3), 432-9 (2013).

[6] Marks, L. B., Yorke, E. D., Jackson, A., Ten Haken, R. K., Constine, L. S., Eisbruch, A., Bentzen, S. M., Nam, J., and Deasy, J. O., "Use of normal tissue complication probability models in the clinic," Int J Radiat Oncol Biol Phys 76(3 Suppl), S10-9 (2010).

[7] Low, D. A., Harms, W. B., Mutic, S., and Purdy, J. A., "A technique for the quantitative evaluation of dose distributions," Med Phys 25(5), 656-61 (1998).

[8] Worthy, D. and Wu, Q., "Dosimetric assessment of rigid setup error by CBCT for HN-IMRT," J Appl Clin Med Phys 11(3), 3187 (2010).

[9] Schwartz, D. L., Garden, A. S., Thomas, J., Chen, Y., Zhang, Y., Lewin, J., Chambers, M. S., and Dong, L., "Adaptive radiotherapy for head-and-neck cancer: initial clinical outcomes from a prospective trial," Int J Radiat Oncol Biol Phys 83(3), 986-93 (2012).

[10] Castadot, P., Geets, X., Lee, J. A., Christian, N., and Gregoire, V., "Assessment by a deformable registration method of the volumetric and positional changes of target

volumes and organs at risk in pharyngo-laryngeal tumors treated with concomitant chemo-radiation," Radiother Oncol 95(2), 209-17 (2010).

[11] Castadot, P., Lee, J. A., Geets, X., and Gregoire, V., "Adaptive radiotherapy of head and neck cancer," Semin Radiat Oncol 20(2), 84-93 (2010).

[12] Nishi, T., Nishimura, Y., Shibata, T., Tamura, M., Nishigaito, N., and Okumura, M., "Volume and dosimetric changes and initial clinical experience of a two-step adaptive intensity modulated radiation therapy (IMRT) scheme for head and neck cancer," Radiother Oncol 106(1), 85-9 (2013).

[13] Chen, A. M., Daly, M. E., Cui, J., Mathai, M., Benedict, S., and Purdy, J. A., "Clinical outcomes among patients with head and neck cancer treated by intensity-modulated radiotherapy with and without adaptive replanning," Head Neck 36(11), 1541-6 (2014).

[14] Neylon, J., Qi, X., Sheng, K., Staton, R., Pukala, J., Manon, R., Low, D. A., Kupelian, P., and Santhanam, A., "A GPU based high-resolution multilevel biomechanical head and neck model for validating deformable image registration," Med Phys 42(1), 232-43 (2015).

[15] Min, Y., Neylon, J., Shah, A., Meeks, S., Lee, P., Kupelian, P., and Santhanam, A. P., "4D-CT Lung registration using anatomy-based multi-level multi-resolution optical flow analysis and thin-plate splines," Int J Comput Assist Radiol Surg 9(5), 875-89 (2014).

[16] Ahn, P. H., Chen, C. C., Ahn, A. I., Hong, L., Scripes, P. G., Shen, J., Lee, C. C., Miller, E., Kalnicki, S., and Garg, M. K., "Adaptive planning in intensity-modulated radiation therapy for head and neck cancers: single-institution experience and clinical implications," Int J Radiat Oncol Biol Phys 80(3), 677-85 (2011).

[17] Smyth, G., McCallum, H. M., Lambert, E. L., and Lawrence, G. P., "A dose distribution overlay technique for image guidance during prostate radiotherapy," Br J Radiol 81(971), 890-6 (2008).

[18] Sharma, M., Weiss, E., and Siebers, J. V., "Dose deformation-invariance in adaptive prostate radiation therapy: implication for treatment simulations," Radiother Oncol 105(2), 207-13 (2012).

[19] Pukala, J., Gray, T., Meeks, S., Manon, R., and Staton, R., "Adaptive radiation therapy replanning for head-and-neck cancers and the dosimetric benefit to the parotid glands," Internation Journal of Radiation Oncology Biology Physics 87, S713-S714 (2013).

[20] Pukala, J., Staton, R., and Langen, K., "What is the importance of dose recalculation for adaptive radiotherapy dose assessment?," Medical Physics 39 (2012).

[21] Hong, T. S., Tome, W. A., Chappell, R. J., Chinnaiyan, P., Mehta, M. P., and Harari, P. M., "The impact of daily setup variations on head-and-neck intensity-modulated radiation therapy," Int J Radiat Oncol Biol Phys 61(3), 779-88 (2005).

[22] National Cancer Institute: Head and neck cancers. Available from: http://www.cancer.gov/cancertopics/factsheet/Sites-Types/head-and-neck. Accessed: August 20, 2013.

CHAPTER 5: Feasibility of Margin Reduction for Level II and III Planning Target Volume in Head-and-Neck Image-Guided Radiotherapy – Dosimetric Assessment via A Deformable Image Registration Framework

A version of this chapter has been published as a manuscript in Current Cancer Therapy Reviews, Vol. 10, No. 4, pp.323-333, 2014. doi: 10.2174/1573394710041504071302229

ABSTRACT

Purpose. To improve normal tissue sparing for head-and-neck (H&N) image-guided radiotherapy (IGRT) by employing treatment plans with tighter margins for CTV 2 and 3, and documenting the delivered dose throughout the entire treatment course.

Methods. Ten H&N cases treated with simultaneous integrated boost on a TomoTherapy unit (Accuray Inc.) were analyzed. Dose-limiting critical structures included brainstem, spinal cord, cochleae, parotid glands and mandible. The targets include the PTV1 (gross disease volume), PTV2 (next echelon nodal regions) and PTV3 (areas harboring subclinical disease). The standard margin plans (plan_ref) were generated using the standard margin of 3 mm to CTV1-3. Reduced margin plans (plan_0margin) using the CTV-to-PTV margin of zero for CTV2 and 3 were compared with plan_ref. All patients went through daily pre-treatment megavoltage CT (MVCT) and weekly kilovoltage CT (kVCT) scans. A GPU-based 3D image deformation/visualization tool was developed to register the weekly kVCT scans with the planning CT scan. The deformation of each contoured structures was computed to account for non-rigid change in the patient setup. Calculation of the dose accumulation was performed to determine the delivered mean/minimum/maximum dose, dose volume histograms (DVHs), etc.

Results. The averaged planned cord maximum doses in Plan_0margin were 7.6% lower, and the parotid mean doses were 18.9% lower than plan_Ref. No significant changes in

D_{95} and D_{90} for the CTV2/3 cumulative doses in both reference and Plan_0margin were observed during the planning stage. Under kVCT guidance on TomoTherapy, for the reference plans, the averaged cumulative mean dose ratios during the entire treatment course were consistent within 5% and 1.5% of the planned mean doses for PTVs and CTVs, respectively. Interfraction anatomical changes introduced variations in delivered target doses that reduced the improved normal structure sparing observed in plan_0margin during the planning stage. For the tighter margin plans, the cumulative mean dose ratios were consistent within 4.3% and 2.3% of the planned mean doses for CTV2 and CTV3, respectively. Similar dose variations of the delivered dose were seen for the reference and tighter margin plans. However, the delivered maximum and mean doses for the cord were 20% and 10% higher than the planned doses; a 3.6% higher cumulative mean dose for the parotids was also observed for the delivered dose than the planned doses in both plans.

Conclusion. The GPU-based image framework enables real-time dose verification, accumulation and documentation. By imposing tighter CTV margins for level 2 and 3 targets for H&N irradiation, acceptable cumulative doses were achievable when coupled with weekly kVCT guidance while improving normal structure sparing.

INTRODUCTION

Radiotherapy has been an effective form to treat head and neck cancer (H&N) in conjunction of chemotherapy. For H&N patients, tumors can be located in the paranasal sinuses, nasal cavity, oral cavity, pharynx and larynx.

Since the head and neck region includes critical structures, the main concern of the treatment is not only an increased survival rate but also protecting the function of these organs [1-5]. State-of-the-art advancements in conformal radiotherapy enabled highly conformal dose distributions that protects the critical structures such as the spinal cord and the parotid glands with the availability of improved dose distributions [6-11].

Undetected and uncompensated factors such as patient posture changes from one treatment fraction to another, and physiological changes such as weight loss or tumor regression may ultimately affect the delivered dose [12-14]. With the advent of image guidance technology, real time imaging was coupled with conformal radiotherapy to form a key tool for quantifying such undetected and uncompensated factors [15-20]. Physicians are able to deliver a planned dose to target more accurately while sparing normal healthy tissue by reducing margins [21-23]. In standard conformal radiation therapy, a 3 to 5 mm margin is given to all the PTVs to compensate for set-up error. However, these safety margins cause an increase in the volume of the high dose region. Since the distance between critical structure and planning target volume decreases during the treatment course, OARs can enter the high dose region. As a result, these organs receive a higher dose than planned [24, 25]. Even though a suitable margin has a small effect on dose volume histogram (DVH) and equivalent uniform dose (EUD), tighter treatment margins are necessary when a tumor touches critical structures. It is especially important when such geometric error occurs [12].

In this paper, we performed a study to investigate the feasibility of developing treatment plans with tighter margins to CTV2 and CTV3 as a way to minimize critical structure dose. Specifically, we investigated the feasibility of a 0 mm margin IMRT plan for head and neck tumors. We focused on the normal organ dose at both the planning stage and the delivery stage where patient specific geometric changes occur. The variations in patient geometry were incorporated using a weekly kilovoltage CT imaging. The delivered dose for the 0 mm margin treatment plan was compared with a standard 3 mm margin treatment plan to quantify the amount of critical structure dose that was minimized during the planning and the delivery stages.

MATERIALS and METHODS

Patient Characteristics

Ten head and neck cancer patients treated with a simultaneous integrated boost IMRT technique on a TomoTherapy unit (Accuray Inc., Sunnyvale, CA) were considered in this work. Table 5.1 shows patient characteristics for the patients included in this study. All patients received daily pretreatment MVCT scans and weekly kVCT scans during the course of treatment. The planning kVCT images were acquired on a Philips Brilliance CT system (Philips Medical Systems, Best, The Netherlands). All patient kVCT images were acquired with the patient in the simulated treatment position with a 50-70 cm FOV, 512x512 inplane resolution, and a 3 mm slice thickness. In total, 71 weekly kVCT scans were analyzed. The patients' weight was recorded weekly during the treatment course.

IMRT Treatment Planning on TomoTherapy

The clinical tumor volumes (CTVs) were delineated on the planning CT by adhering to the principle of respecting anatomic boundaries. CTV1 was defined as any visible tumor mass as delineated on imaging studies, whether at the primary site or cervical lymphatics. It often

Table 5.1. Patient characteristics.

Pt #	Diagnosis	Prescription (Gy)	No. of fx	Dose/fx (Gy)	Initial Weight (kg)	Weight Change (kg)	Weight Change (%)
1	Tonsil	70	35	2	72.1	-8.8	-12.2
2	BOT	70	35	2	63.5	-0.9	-1.4
3	Nasopharynx	69.96	33	2.12	81.6	-5.4	-6.6
4	Tonsil	70	35	2	95.7	-10.9	-11.4
5	Tonsil	70	35	2	83.9	-7	-8.3
6	Tonsil	70	35	2	88	-2.1	-2.4
7	Tonsil	70	35	2	93.4	-9.6	-10.3
8	Tonsil	70	35	2	107.5	-11.5	-10.7
9	Tonsil	66	30	2.2	86.4	-4.4	-5.1
10	BOT	70	35	2	99.8	-10.7	-10.7

Abbreviation: BOT: Base of tongue; fx = fraction

coincided with the gross tumor volume (GTV) plus the perceived direct disease extension, and may encompass the entire anatomic structure (such as the nasopharynx) to which the treating radiation oncologist feels necessary to deliver tumoricidal dosage sufficient for controlling a bulky tumor (traditionally held to be around 70 Gy in 2-Gy per fraction scheme). CTV2 was defined as either an adjacent area or structure perceived to be at risk, or the next echelon lymphatic drainage areas. For post-resection cases, it also included surgical bed where a somewhat moderate level of dosage (*e.g.* 60 Gy) may be needed in order to compensate for the perceived accelerated repopulation of residual tumor cells. Finally, CTV3 is defined as any target volume which may harbor only subclinical (*i.e.* undetectable clinically) disease such as micro-metastases, for which a relatively low dose level (*e.g.* 50 Gy) might be sufficient. In general, these CTV structures are determined based on each individual physician's practicing philosophy with respect to the tumor's perceived anatomic extent. The critical structures including brainstem, spinal cord, cochleae, parotids, mandible, etc were also delineated. Two IMRT plans were created using different CTV-to-PTV margin. For the standard plan, a CTV-to-

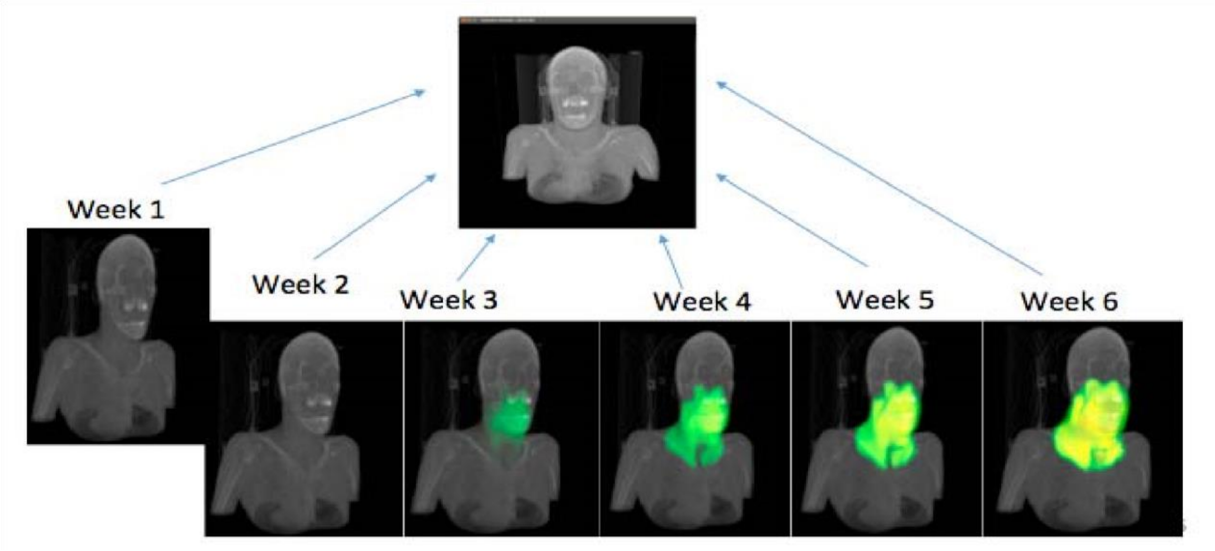


Figure 5.1. A schematic illustration of the dose accumulation using the weekly CT scans.

PTV margin of 3 mm was given for PTV1-3 (plan_ref); a reduced margin plan (plan_0margin) was created using 3 mm margin for PTV1 while zero margins were employed for CTV2 and 3. The prescription doses were in the range of 50-70 Gy in 30-36 fractions for PTV1 (gross disease volume), PTV2 (next echelon nodal regions) and PTV3 (areas harboring subclinical disease). Before each treatment, alignment was performed using in-room lasers and 3-point markers on the patient with the customized immobilization device.

For all cases, all treatment plans were created on the TomoTherapy planning system (version 4.0) using the following parameters: field width of 2.5 cm, modulation factor of 2.5 and pitch of 0.287.

In-House Deformable Image Registration (DIR) Framework

Tracking anatomical changes is crucial to account for geometric changes in the patient anatomy [26, 27]. We employed an in-house GPU-based dense optical flow registration algorithm for

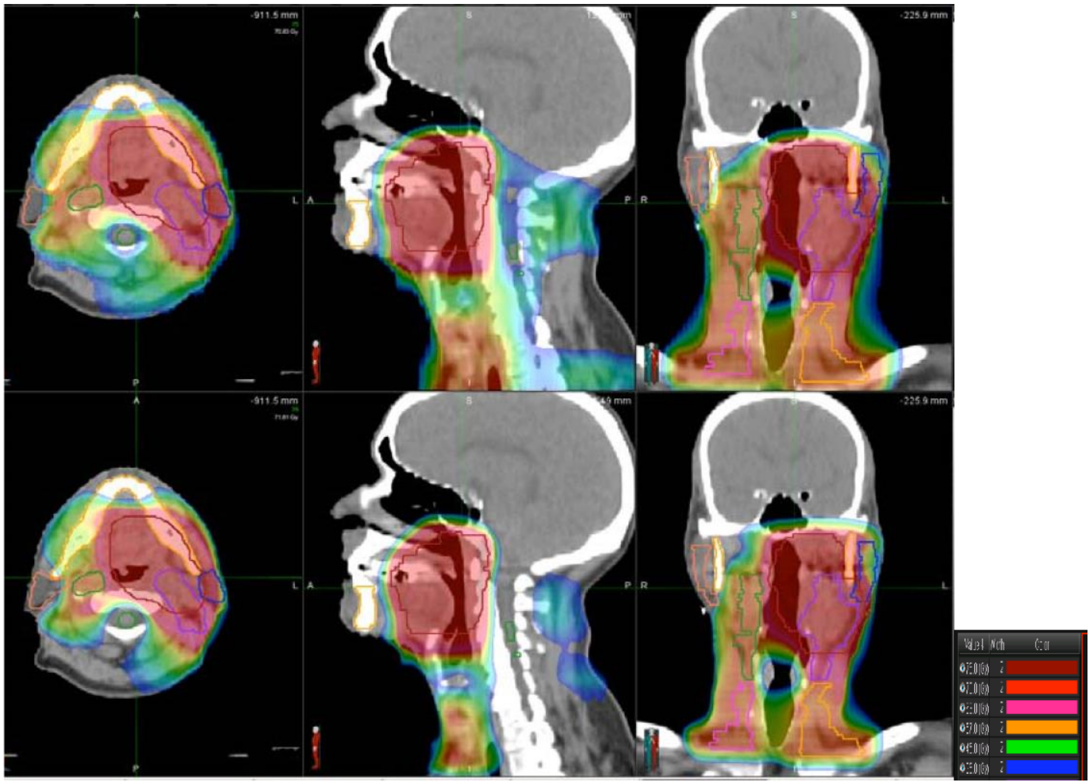


Figure 5.2. Dose distribution on a transverse, sagittal and coronal view of the standard margin plan and the reduced margin plan for a representative case (patient #5).

registering planning kVCT with the weekly kVCT scan [28]. We considered the weekly kVCT scan as the target 3D image and the planning kVCT as the source 3D image.

The DICOM objects for each patient, including treatment planning CT, planning CT structure set, planning dose and the weekly kVCT images were exported to an in-house DIR framework.

As a first step, the source/target pair was resampled to have the same image dimensions and resolutions. A multi-resolution registration approach was used to account for voxel displacement greater than 1 voxel distance. The number of resolution levels and the smoothness values were set to 5 and 150 as they provided optimal registration to account for non-rigid geometric continuity. The registration process computed the displacement vectors associated with each voxel in the planning kVCT scan. The treatment plan corresponding to the planning kVCT scan was finally warped to compute the dose to be delivered that corresponded with the weekly kVCT scan. Finally, the doses to be delivered to critical structures were recomputed.

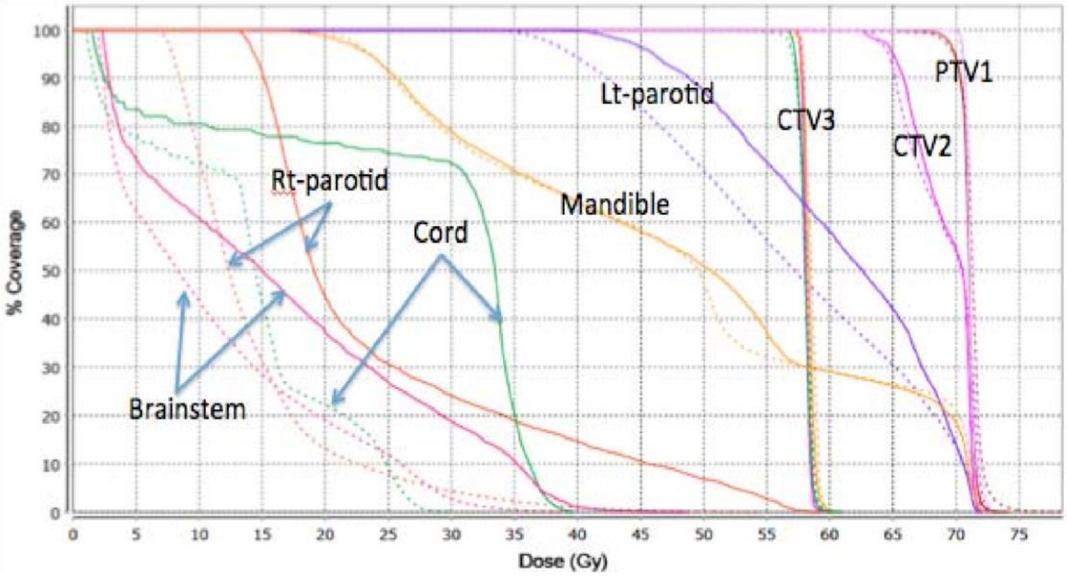


Figure 5.3. DVH comparison of the standard plan (solid line) versus the reduced margin plan (dotted line) for patient 5.

Figure 5.1 shows a schematic illustration of the dose accumulation using the weekly CT scan. The dose to be delivered. It can be seen that at the end of the treatment fractions, the dose delivered for the critical structures and the tumor was documented for each voxel.

Dose Accumulation for the Reduced Margin Plan

To simulate the delivered dose and cumulative dose for the reduced margin plan, we used the weekly kVCT scans acquired for the standard margin plan at treatment position. Each pretreatment weekly kVCT was registered using our in-house DIR framework and deformed to the corresponding planning CT scan. The deformed new structure set (with zero expansions of CTV-to-PTV for level II and III) representing the anatomy on a given treatment fraction populated from the planning CT. The delivered dose distributions based on plan_0margin for the targets and critical structures were computed and compared with the plan_ref.

RESULTS

In this section, we first present our results on comparing the treatment plans developed with zero margins for CTV2 and CTV3 with the treatment plans developed using conventional margins. We then present our results on comparing the delivered dose for both the CTVs and the critical structures using the two planning strategies. Ten head-and-neck cases were analyzed and presented in this work. The DVHs of the standard plan (plan_ref) and the reduced margin plan (plan_0margin), and the actual accumulated doses for both plans were calculated and compared.

Dosimetric Comparison of the Standard Plan Versus the Reduced Margin Plan

Figure 5.2 shows the dose distribution of the standard margin plan (top) versus the reduced margin plan (bottom) for a representative case (patient 5). The DVHs of the selected structures for the same case are displayed in figure 5.3. For both the standard plan and the reduced margin plan, the PTV1 remains sufficient coverage, the CTV2 and CTV3 shows no significant difference, while great OAR sparing for the cord, the brainstem, left- and right- parotid glands are clearly seen in the reduced margin plan. The detailed planned dose metrics, such as maximum cord dose, mean doses of the left- and right- parotid glands, the maximum and average doses for the PTV1, CTV2 and CTV3 of both standard and the reduced margin plans are tabulated in Table 5.2.

Figure 5.4 shows the ratios of the selected dosimetric parameters for the reduced margin plans and the standard margin plans. It appears that the mean doses for CTV2 and 3 are consistent within 3% and 4.9% respectively between the plans with and without margin. However, large variations (up to 45%) of parotid gland mean dose sparing was seen for patient #1 and #5 for zero margin plan; up to 30% of the cord max dose was observed compared to the standard

Table 5.2. Dosimetric Parameters Between the Standard Margin Versus the Reduced Margin Plan.

	PTV1		CTV2		CTV3		Cord	Lt-Parotid	Rt-Parotid
	Max Dose (Gy)	Ave Dose (Gy)	Max Dose (Gy)	Ave Dose (Gy)	Max Dose (Gy)	Ave Dose (Gy)	Max Dose (Gy)	Ave Dose (Gy)	Ave Dose (Gy)
Pt #	**Standard Margin Plan**								
1	76.69	71.83	74.43	72.25	74.0	72.29	41.51	22.4	38.88
2	74.39	71.01	73.36	71.82	72.4	67.67	41.83	24.02	61.41
3	74.45	71.37	73.55	66.05	72.9	66.99	33.22	56.42	49.19
4	74.22	72.07	73.7	70.39	62.36	57.71	44.91	62.98	23.37
5	74.11	70.99	72.92	71.07	72.68	69.22	39.65	60.93	24.55
6	77.93	71.13	75.03	68.02	70.90	64.23	39.63	24.85	24.56
7	73.99	71.43	73.99	71.7	73.48	67.7	43.39	40.98	14.02
8	76.47	71.56	76.47	69.05	73.3	59.13	44.32	27.48	22.92
9	69.33	67.31	68.93	66.61	68.93	58.91	42.71	7.56	38.83
10	74.16	71.14	73.01	67.02	72.12	61.99	40.18	22.76	20.81
Mean	74.54	70.97	73.51	69.62	71.12	65.44	41.00	29.84	29.06
Pt #	**Reduced Margin Plan**								
1	74.65	71.68	72.46	70.11	71.78	68.75	38.11	19.55	29.35
2	74.61	71.35	72.80	69.79	73.06	67.15	29.06	13.09	59.03
3	73.57	69.85	72.81	67.06	71.27	65.16	27.77	51.34	48.92
4	75.45	72.48	75.06	70.38	61.59	57.81	34.7	58.71	19.37
5	78.26	71.27	72.03	68.98	73.25	69.2	29.35	57.01	14.14
6	76.38	70.9	77.55	67.79	69.83	63.91	36.73	21.03	19.90
7	76.21	71.35	73.96	71.24	72.35	66.81	38.68	40.83	12.72
8	77.0	71.86	76.76	69.11	73.96	59.1	41.24	26.52	21.16
9	71.6	67.84	70.31	66.77	69.95	59.15	43.89	6.87	37.68
10	73.87	71.43	73.48	66.93	72.56	62.04	39.53	21.24	12.85
Mean	75.13	70.99	73.69	68.79	70.87	63.78	35.50	26.08	23.87

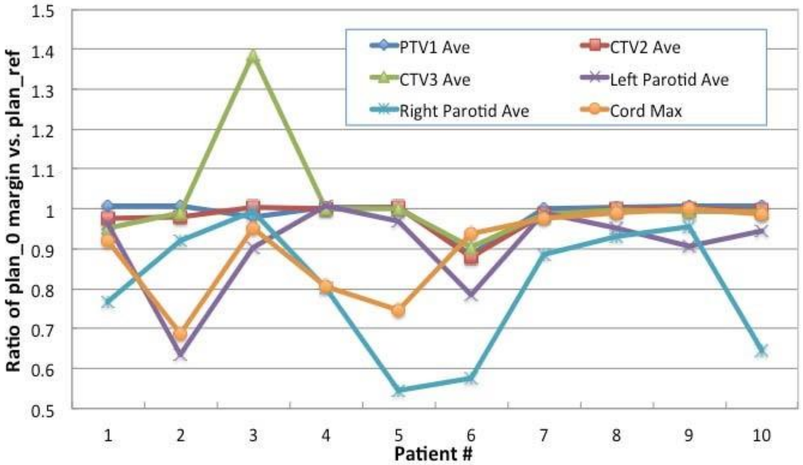


Figure 5.4. The ratios of the dosimetric metrics between the standard margin plan and the reduced margin plan for the targets and the selected structures.

margin plan. Student t-test was performed for the dosimetric parameters between the standard margin versus reduced margin plans, the p-values are 0.01, 0.40, and 0.38 for the maximum cord, mean left-parotid and right parotid glands, respectively. For the targets, comparable doses were found for all PTV1, CTV2 and CTV3. Such observations support the fact that treatment plans with zero margins for CTV2 and CTV3 facilitate a treatment that delivers the same dose to the tumor volume as that of a conventional treatment plan while dramatically reducing the dose delivered to organ-at-risks.

Delivered Cumulative Dose Comparison

Each pre-treatment image was acquired to ensure the correct patient alignment and thus the delivered dose distribution to match with the planned dose distribution. Figure 5.5 displays the comparison of the actual delivered dose for the parotid glands between the standard margin plan and the reduced margin plans for the group of 10 cases. The delivered dose, in general, agreed well with the planned doses for both standard margin and reduced margin plans. For the cord maximum dose, left-parotid mean dose and right-parotid mean dose, the planned dose

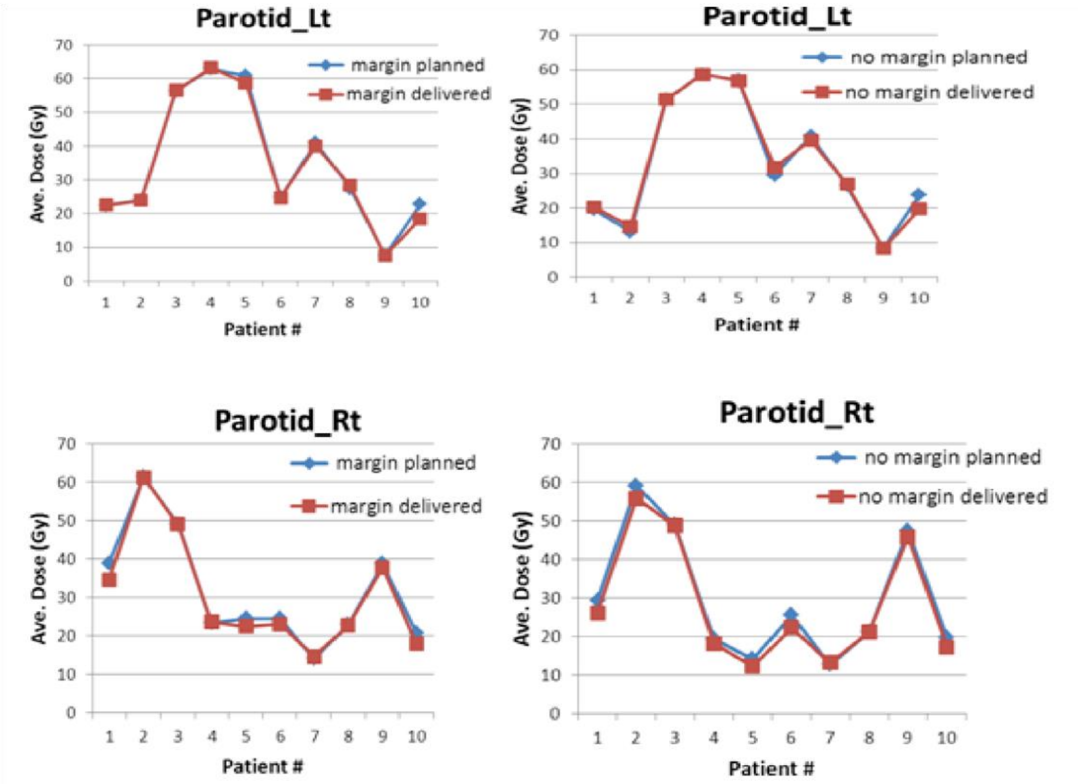
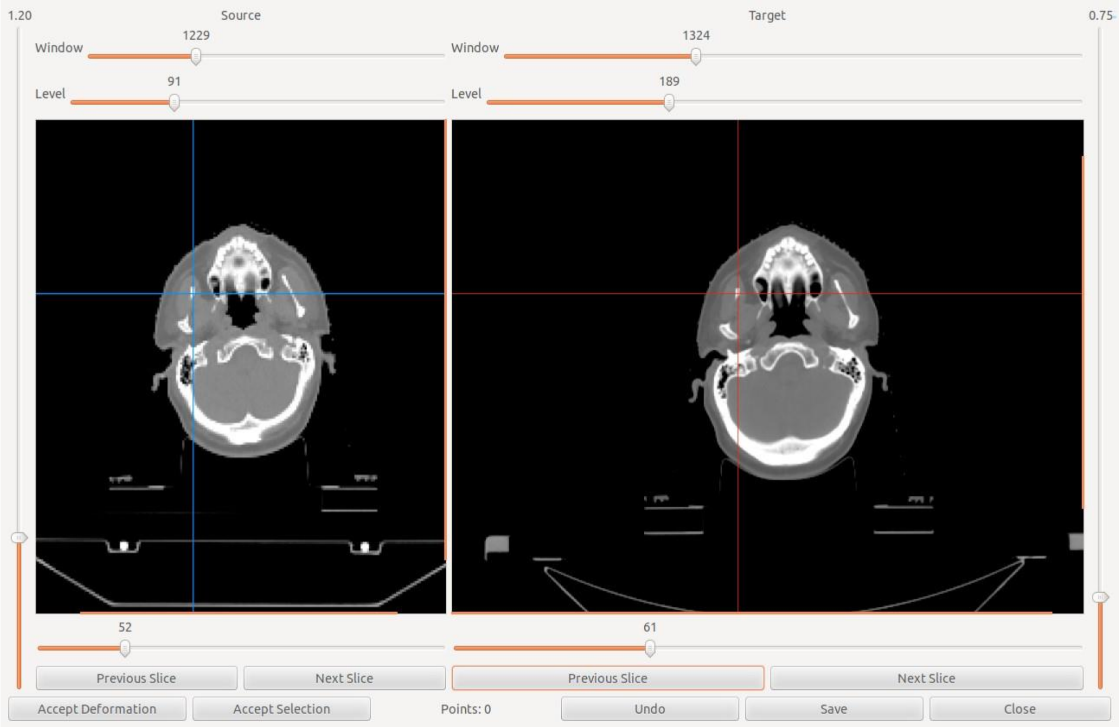


Figure 5.5. Comparison of the actual delivered dose for the parotid glands between the standard margin plan and the reduced margin plans.

versus the delivered dose are p=0.19 (the standard margin) *vs.* 0.31 (the tighter margin); p=0.45 (the standard margin) *vs.* 0.44 (the reduced margin); p=0.43 (the standard margin) *vs.* 044 (the reduced margin), respectively. Between the delivered doses with the standard margin versus no margin, the p=0.16, 0.45 and 0.49 respectively. By the end of the treatment course, all clinical target volumes received the acceptable doses as was expected.

The accumulated dose was checked for all patients to evaluate how closely the planned dose distribution was followed. Since kilovoltage computed tomography was used for dose calculation, each weekly dose was corrected based on a fraction in order to compare the two following weeks. Based on this comparison. The weekly accumulated dose given to the patients is similar equivalent to the weekly planned dose. Moreover, the total delivered dose is the same as with the planned dose. As a result, tumor coverage was provided with the zero margin plans.



1.20 Source

Window 1229

Level 91

Target 0.75

Window 1324

Level 189

52

61

Points: 0

Figure 5.6. The user interface developed for performing a landmark based registration validation.

Validation of the In-House Deformation Framework

The accuracy and the robustness of the analysis were greatly dependent on the accuracy of the in-house registration algorithm [27]. We validated the head and neck registration using a landmark-based Target Registration Error (TRE) metric [29]. For our analysis, we considered the planning kVCT to be source 3D image and the kVCT of the last week of treatment to be the target 3D image. A set of 80 landmarks was marked on the rigid structures of a reference kVCT and tracked from one kVCT

dataset to another. Figure 5.6 presents the user interface that we employed for our validation process. The landmarks were placed by an expert on the reference kVCT data (left) as shown by the cross hairs. For each of the landmarks, the corresponding landmark in the target kVCT data was calculated using the image registration results and were visually shown to the expert as cross hairs overlapping the target (right) image. Based on the results, the expert either accepted the registration results or marked the correct landmark on the target image. Once the 80 landmarks were delivered, the TRE for each of the datasets were computed. Table 5.3 tabulates the DIR registration results for the group of ten cases. For each of the datasets, the TRE was found to be in the range of 0.5-1.13 mm.

Table 5.3. Landmark-based validation of the in-house registration algorithm.

Pt	1	2	3	4	5	6	7	8	9	10
TRE (mm)	0.72	0.65	0.94	1.3	1.4	1.1	0.9	0.8	0.9	0.7

DISCUSSION

One of the major challenges of the IMRT treatment is to minimize doses to the critical structures while providing the intended dose to the target. To ensure sufficient target coverage, the common practice is imposing a proper margin to the planning target volume from the clinical tumor volume to account for uncertainties in planning or treatment delivery [30]. The PTV is meant to encompass beyond CTVs by compensating for patient set-up and motion uncertainties, which may take on a random (Gaussian) orientation 3dimensionally. Such CTV-to-PTV margin is thus driven by the practicing physics protocol, and may be modified based on the existing patient setup motion compensation technology such as image guided radiation therapy (IGRT). With IGRT, we aim to better spare the OARs by further reduction in such CTV-to-PTV margins, in particular for CTV2 to PTV2 and CTV3 to PTV3 expansions. As for PTV1, we feel that adequate margins beyond CTV1 should still be preserved despite image-guidance endeavor, since the outlining of CTV1s by physicians already entails certain degree of educated guess such that any systemic reduction of the perceived gross tumor extent by physics protocol could translate into significant compromise in the ultimate tumor control probability. It is known that PTV is a geometric concept that takes into consideration the net effect of all possible geometric variations and is used to ensure that the CTV receives the prescribed dose.

These geometric uncertainties include organ delineation, setup errors, and organ motion that occur throughout the planning and treatment process. The clinical implementation of margin reduction may also depend on the patient immobilization devices, the image quality [31], IGRT procedures, the anatomic sites, etc. [32].

Stroom *et al.* [33] and van Herk *et al.* [34] derived CTV-to-PTV margin recipes accounting for the systematic and random setup errors. An important shortcoming of these margin recipes is their lack of adequately incorporating both rotational and morphologic errors [35]. In this study, we used a novel approach to assess the feasibility of margin reduction *via* a GPU-based framework. Through a retrospective study of real IGRT images, we simulated the daily and cumulative dose distribution if reduced margins for CTV II and III were imposed when patients were in the actual treatment position. With image guidance, the dose distribution based on the reduced margin plans appeared to be acceptable for the CTV2 and 3; in the meantime, better OAR sparing, compared with the standard margin plan, would be possible. For the patients who had large anatomic changes, such as target shrinkage, weight loss, etc., during the course of the treatment, an adaptive plan based on the new anatomic could be considered.

The reproducibility of patient setup is of particular importance for head and neck IMRT treatment due to the proximity of targets to the critical structures and the sharp dose falloff of the planned dose distribution. The standard patient immobilization device for head and neck irradiation is a customized head and neck thermoplastic masks. However, the head and neck mask may not provide sufficient immobilization of the shoulders, which is of importance in comprehensive nodal irradiation in the neck area. The reduction of CTV2 and CTV3 margin calls for better patient immobilization devices, such as the head and neck shoulder mask or better robust patient alignment procedure, to provide better immobilization of the entire upper part of the body in the treatment position. Clinical validation is needed to verify the immobilization accuracy of the device. Caution needs to be taken when tightening the CTV-to-PTV margin in clinical practice.

In addition, the CTV-to-PTV margin of head-and-neck cancer may be affected by the imaging modality. Various IGRT modalities, such as kV cone beam CT (kVCBCT), mega voltage cone beam CT (MVCBCT), mega voltage fan beam CT (MVCT) are available and widely used in clinical practice. The image quality obtained from these on-board CT systems is not as good as the planning kVCT. As a result, large margin may be necessary for the on-board image systems with inferior image quality due to large random error [31].

CONCLUSION

We presented a feasibility study of potential margin reduction for Level II and III planning target volumes in image-guided H&N radiotherapy. An in-house GPU-based deformable image registration framework was used to compute the delivered dose based on weekly images and the delivered accumulative dose during the entire course. Reduce CTV-to-PTV expansion for level II and III targets for H&N irradiation may greatly reduce the dose delivered to the critical structures, such as the parotid glands and cord. However, it was observed that subject-specific anatomical changes led to a higher dose delivered to critical structures. Thus, while using tighter margins for the CTV2 and CTV3 may lead to better sparing of normal tissues, adaptive re-planning will be required in order to account for changes in the patient geometry. The kVCT guidance with zero CTV-to-PTV margin appears to result in acceptable cumulative doses to the targets (CTV2 and CTV3) while greatly improving normal structure sparing.

Future work would focus on developing adaptive radiotherapy strategies for head and neck radiotherapy that will ensure zero margin treatment plans are delivered accounting for changes in the patient geometry. Advancements in image registration and biomechanical head and neck modeling will lead to a precise tracking of patient anatomy changes from one treatment fraction to another. Such adaptive radiotherapy strategies will eventually lead to a better sparing of normal organs and to a better patient quality of life. Future work would also include a systematic analysis of random errors that will have to be included in the zero-margin treatment plans and their impact on the dose improvements. Such a study would be critical to document the need the for algorithm improvements in image registration and biomechanical modeling as a way to minimize the impact of random errors on the dosimetric improvements provided by the zero-margin treatment plans.

REFERENCES

[1] Argiris, A., Karamouzis, M. V., Raben, D., and Ferris, R. L., "Head and neck cancer," Lancet 371(9625), 1695-709 (2008).

[2] Yom, S. S., Raben, D., Siddiqui, F., Lu, J. J., and Yao, M., "Skull Base Head and Neck Cancer," in [Stereotactic Body Radiation Therapy], Springer Berlin Heidelberg, 267-284 (2012).

[3] McMahon, S. and Chen, A. Y., "Head and neck cancer," Cancer Metastasis Rev 22(1), 21-4 (2003).

[4] Blanco, A. I., Chao, K. S., El Naqa, I., Franklin, G. E., Zakarian, K., Vicic, M., and Deasy, J. O., "Dose-volume modeling of salivary function in patients with head-and-neck cancer receiving radiotherapy," Int J Radiat Oncol Biol Phys 62(4), 1055-69 (2005).

[5] Chambers, M. S., Garden, A. S., Kies, M. S., and Martin, J. W., "Radiation-induced xerostomia in patients with head and neck cancer: pathogenesis, impact on quality of life, and management," Head Neck 26(9), 796-807 (2004).

[6] Dietrich, S., Rodgers, J., and Chan, R., "Radiosurgery," in [Image-Guided Interventions], Springer US, 461-500 (2008).

[7] Eshleman, J. S., "A New Tool to Help Fight Cancer - Tomotherapy," Journal of Lancaster General Hospital 4(3), 106-112 (2009).

[8] Ploquin, N. P., Belec, J. G., and Clark, B. G. "Dosimetric Comparison between Helical Tomotherapy and Biologically Based IMRT Treatment Planning System for Selected Cases," in *World Congress on Medical Physics and Biomedical Engineering*, Munich, Germany: Springer Berlin Heidelberg (2009).

[9] Hong, T. S., Tome, W. A., and Harari, P. M., "Intensity-modulated radiation therapy in the management of head and neck cancer," Curr Opin Oncol 17(3), 231-5 (2005).

[10] Chen, A. M., Jennelle, R. L., Sreeraman, R., Yang, C. C., Liu, T., Vijayakumar, S., and Purdy, J. A., "Initial clinical experience with helical tomotherapy for head and neck cancer," Head Neck 31(12), 1571-8 (2009).

[11] Lee, N., Xia, P., Fischbein, N. J., Akazawa, P., Akazawa, C., and Quivey, J. M., "Intensity-modulated radiation therapy for head-and-neck cancer: the UCSF experience focusing on target volume delineation," Int J Radiat Oncol Biol Phys 57(1), 49-60 (2003).

[12] Lee, C., Langen, K. M., Lu, W., Haimerl, J., Schnarr, E., Ruchala, K. J., Olivera, G. H., Meeks, S. L., Kupelian, P. A., Shellenberger, T. D., and Manon, R. R., "Assessment of parotid gland dose changes during head and neck cancer radiotherapy using daily megavoltage computed tomography and deformable image registration," Int J Radiat Oncol Biol Phys 71(5), 1563-71 (2008).

[13] Toledano, I., Graff, P., Serre, A., Boisselier, P., Bensadoun, R. J., Ortholan, C., Pommier, P., Racadot, S., Calais, G., Alfonsi, M., Favrel, V., Giraud, P., and Lapeyre, M., "Intensity-modulated radiotherapy in head and neck cancer: results of the prospective study GORTEC 2004-03," Radiother Oncol 103(1), 57-62 (2012).

[14] van Vulpen, M., Field, C., Raaijmakers, C. P., Parliament, M. B., Terhaard, C. H., MacKenzie, M. A., Scrimger, R., Lagendijk, J. J., and Fallone, B. G., "Comparing step-and-shoot IMRT with dynamic helical tomotherapy IMRT plans for head-and-neck cancer," Int J Radiat Oncol Biol Phys 62(5), 1535-9 (2005).

[15] Beavis, A. W., "Is tomotherapy the future of IMRT?," Br J Radiol 77(916), 285-95 (2004).

[16] Dawson, L. A. and Jaffray, D. A., "Advances in image-guided radiation therapy," J Clin Oncol 25(8), 938-46 (2007).

[17] William, B. W., "Image Guided Radiotherapy," 5, 86-92 (2008).

[18] Chung, Y., Yoon, H. I., Kim, J. H., Nam, K. C., and Koom, W. S., "Is helical tomotherapy accurate and safe enough for spine stereotactic body radiotherapy," Journal of Cancer Research and Clinical Oncology 139(2) (2012).

[19] Dawson, L. A. and Sharpe, M. B., "Image-guided radiotherapy: rationale, benefits, and limitations," Lancet Oncol 7(10), 848-58 (2006).

[20] Graff, P., Hu, W., Yom, S. S., and Pouliot, J., "Does IGRT ensure target dose coverage of head and neck IMRT patients?," Radiother Oncol 104(1), 83-90 (2012).

[21] Schwarz, M., Giske, K., Stoll, A., Nill, S., Huber, P. E., Debus, J., Bendl, R., and Stoiber, E. M., "IGRT versus non-IGRT for postoperative head-and-neck IMRT patients: dosimetric consequences arising from a PTV margin reduction," Radiat Oncol 7, 133 (2012).

[22] Verellen, D., De Ridder, M., Linthout, N., Tournel, K., Soete, G., and Storme, G., "Innovations in image-guided radiotherapy," Nat Rev Cancer 7(12), 949-60 (2007).

[23] Hong, T. S., Tome, W. A., Chappell, R. J., Chinnaiyan, P., Mehta, M. P., and Harari, P. M., "The impact of daily setup variations on head-and-neck intensity-modulated radiation therapy," Int J Radiat Oncol Biol Phys 61(3), 779-88 (2005).

[24] Loo, H., Fairfoul, J., Chakrabarti, A., Dean, J. C., Benson, R. J., Jefferies, S. J., and Burnet, N. G., "Tumour shrinkage and contour change during radiotherapy increase the dose to organs at risk but not the target volumes for head and neck cancer patients treated on the TomoTherapy HiArt system," Clin Oncol (R Coll Radiol) 23(1), 40-7 (2011).

[25] *Prescribing, Recording and Reporting Photon Beam Therapy (Supplement to ICRU Report 50), ICRU Report 62*. 1999: ICRU Bethesda, MD. p. ix+52.

[26] Liu, F., Erickson, B., Peng, C., and Li, X. A., "Characterization and management of interfractional anatomic changes for pancreatic cancer radiotherapy," Int J Radiat Oncol Biol Phys 83(3), e423-9 (2012).

[27] Bujold, A., Craig, T., Jaffray, D., and Dawson, L. A., "Image-guided radiotherapy: has it influenced patient outcomes?," Semin Radiat Oncol 22(1), 50-61 (2012).

[28] Lu, W., Olivera, G. H., Chen, Q., Ruchala, K. J., Haimerl, J., Meeks, S. L., Langen, K. M., and Kupelian, P. A., "Deformable registration of the planning image (kVCT) and the daily images (MVCT) for adaptive radiation therapy," Phys Med Biol 51(17), 4357-74 (2006).

[29] Neylon, J., Qi, X., Sheng, K., Staton, R., Pukala, J., Manon, R., Low, D. A., Kupelian, P., and Santhanam, A., "A GPU based high-resolution multilevel biomechanical head and neck model for validating deformable image registration," Med Phys 42(1), 232-43 (2015).

[30] Fitzpatrick, J. M., [Medical Image Regsitration]: CRC Press, (2001).

[31] Qi, X. S., Hu, A. Y., Lee, S. P., Lee, P., DeMarco, J., Li, X. A., Steinberg, M. L., Kupelian, P., and Low, D., "Assessment of interfraction patient setup for head-and-neck cancer intensity modulated radiation therapy using multiple computed tomography-based image guidance," Int J Radiat Oncol Biol Phys 86(3), 432-9 (2013).

[32] Li, X. A., Qi, X. S., Pitterle, M., Kalakota, K., Mueller, K., Erickson, B. A., Wang, D., Schultz, C. J., Firat, S. Y., and Wilson, J. F., "Interfractional variations in patient setup and anatomic change assessed by daily computed tomography," Int J Radiat Oncol Biol Phys 68(2), 581-91 (2007).

[33] Stroom, J. C., de Boer, H. C., Huizenga, H., and Visser, A. G., "Inclusion of geometrical uncertainties in radiotherapy treatment planning by means of coverage probability," Int J Radiat Oncol Biol Phys 43(4), 905-19 (1999).

[34] van Herk, M., Remeijer, P., Rasch, C., and Lebesque, J. V., "The probability of correct target dosage: dose-population histograms for deriving treatment margins in radiotherapy," Int J Radiat Oncol Biol Phys 47(4), 1121-35 (2000).

[35] Meijer, G. J., de Klerk, J., Bzdusek, K., van den Berg, H. A., Janssen, R., Kaus, M. R., Rodrigus, P., and van der Toorn, P. P., "What CTV-to-PTV margins should be applied for prostate irradiation? Four-dimensional quantitative assessment using model-based deformable image registration techniques," Int J Radiat Oncol Biol Phys 72(5), 1416-25 (2008).

CHAPTER 6: A GPU based high-resolution multi-level biomechanical head and neck model for validating deformable image registration

A version of this chapter has been published as a manuscript: Med Phys, Vol. 42, No.1, pp.232-243, 2015. doi: 10.1118/1.4903504

ABSTRACT

Purpose. Validating the usage of deformable image registration (DIR) for daily patient positioning is critical for adaptive radiotherapy applications pertaining to head and neck (HN) radiotherapy. We present a methodology for generating biomechanically realistic ground-truth data for validating DIR algorithms for HN anatomy by (a) developing a high-resolution deformable biomechanical HN model from a planning CT, (b) simulating deformations for a range of inter-fraction posture changes and physiological regression and (c) generating subsequent CT images representing the deformed anatomy.

Methods. The biomechanical model was developed using HN kVCT datasets and the corresponding structure contours. The voxels inside a given 3D contour boundary were clustered using a GPU-based algorithm that accounted for inconsistencies and gaps in the boundary to form a volumetric structure. While the bony anatomy was modeled as rigid body, the muscle and soft tissue structures were modeled as mass-spring-damper (MSD) models with elastic material properties that corresponded to the underlying contoured anatomies. Within a given muscle structure, the voxels were classified using a uniform grid and a normalized mass was assigned to each voxel based on its Hounsfield number.

The soft-tissue deformation for a given skeletal actuation was performed using an implicit Euler integration with each iteration split into two sub-steps: one for the muscle structures, and the other for the remaining soft tissues. Posture changes were simulated by articulating the skeletal structure and enabling the soft structures to deform accordingly. Physiological changes representing tumor regression were simulated by reducing the target volume and enabling the surrounding soft structures to deform accordingly.

Finally, we also discuss a new approach to generate kVCT images representing the deformed anatomy that accounts for gaps and anti-aliasing artifacts that may be caused by the biomechanical deformation process. Accuracy and stability of the model response were validated using ground truth simulations representing soft tissue behavior under local and global deformations. Numerical accuracy of the HN deformations were analyzed by applying non-rigid skeletal transformations acquired from inter-fraction kVCT images to the model's skeletal structures and comparing the subsequent soft tissue deformations of the model with the clinical anatomy.

Results. The GPU based framework enabled the model deformation to be performed at 60 frames per second, facilitating simulations of posture changes and physiological regressions at interactive speeds. The soft tissue response was accurate with an R^2 value of > 0.98 when compared to ground-truth global and local force deformation analysis. The deformation of the HN anatomy by the model agreed with the clinically observed deformations with an average correlation coefficient of 0.956. For a clinically relevant range of posture and physiological changes, the model deformations stabilized with an uncertainty of less than 0.01 mm.

Conclusions. Documenting dose delivery for HN radiotherapy is essential accounting for posture and physiological changes. The biomechanical model discussed in this paper was able to deform in real-time, allowing interactive simulations and visualization of such changes. The model would allow patient specific validations of the DIR method and has the potential to be a significant aid in adaptive radiotherapy techniques.

INTRODUCTION

The term head and neck cancer (HNC) refers to a group of biologically similar cancers originating from the upper aero digestive tract, including the lip, oral cavity (mouth), nasal cavity, Para nasal sinuses, pharynx, and larynx. 90% of head and neck cancers are squamous cell carcinomas (SCCHN), originating from the mucosal lining (epithelium) of these regions [1]. HNC often spread to the lymph nodes of the neck, leading to cancer metastasis in the rest of the patient's body [2]. Radiotherapy (RT) has seen a major push towards treatment plans for the HNC that are tailored to the patient and adapted to their radiation response [3-6]. Ignoring patient mis-alignments caused by non-rigid changes in patient posture and physiology can lead to under-dosing the tumor and over-irradiating the healthy tissue [5, 7]. Image-guided analyses of such non-rigid head and neck anatomy variations were made possible by use of deformable image registration (DIR) frameworks that register the patient planning anatomy with the treatment anatomy. Such analyses have led to several indications on the need for better patient aligning. For instance, Wang et al [8] showed that uncorrected patient positioning misalignments would increase the maximum dose to both the brainstem and spinal cord by 10 Gy and the mean dose to the left and right parotid glands by 7.8 and 8.5 Gy, respectively. Similarly, 95% of the gross tumor volume (GTV) and clinical target volume (CTV) would decrease by 4 Gy and 5.6 Gy, respectively.

The accuracy of DIR to help quantify patient posture and physiological changes is critical for the success of adaptive RT. Adaptive RT will employ quantitative dose delivery error characterization and subsequent compensatory strategies. However, DIR development has been hampered by a lack of techniques that generate ground-truth deformations that can be used to evaluate competing DIR algorithms. This paper focuses on developing a biomechanical model that will be the first step towards generating ground-truth deformations that can be used for validating both image registration and adaptive RT frameworks. Biomechanical human anatomy models have been developed for applications ranging from computer animation to CT image registration.

Sophisticated biomechanical models have been developed for individual anatomical sites, including the head and neck [9], the hand [10-12], lungs [13], and the leg [14]. Such models, when developed from patient CT or MRI, can create subject-specific physiological and musculoskeletal dynamic atlas. As an example, subject specific cardiac models of normal and diseased heart have been developed using Non-Uniform Rational Bezier Splines (NURBS) in order to simulate the cardiac motion before and after the treatment [15]. Physics-based methods, such as Finite Element Methods and Mass-Spring Models, have been applied for deforming anatomy of the torso [16, 17], and the biomechanical nature of these models also allows for the inclusion of subject specific tumor representations and day-to-day variations in the treatment.

The focus of this paper is on the biomechanical head and neck model development that can be used for validating DIR algorithms for the head and neck anatomy. The human head-neck musculoskeletal system is highly complex with approximately 57 articular bones and many more muscle actuators. Comprehensive biomechanical modeling and control of the head and neck anatomy is the most principled approach for simulating subtle deformations such as neck rotation movements and physiological changes such as tumor shrinkage and internal organ movements.

To construct precise ground-truth data for validating DIR, which provides clinically realistic deformations, where the motion of each voxel is known, the biomechanical models need to satisfy the following: (a) the model must have a one-to-one correspondence with the reference anatomy, i.e., for every voxel in the reference anatomy, the model must include a model element (e.g. a node with an assigned mass and elasticity), and vice versa, (b) the model must simulate both posture as well as physiological changes, and (c) the model must be validated to ensure clinical relevance. In this paper, we present a method for deforming a high-resolution biomechanical head and neck model. In order to address the high computational demands of this method, we present algorithms to employ a Graphics Processing Unit (GPU) based computational framework.

The key contribution of this paper are as follows: (a) To our knowledge, this is the first work to demonstrate a GPU framework for deforming a high resolution biomechanical head and neck model at interactive speeds; (b) The model anatomy maintains a one-to-one correspondence between a planning CT anatomy such that the model can compute the displacement of every voxel in the CT without employing any interpolation; and (c) The model also generates an equivalent CT data set for validating image-based registration algorithms.

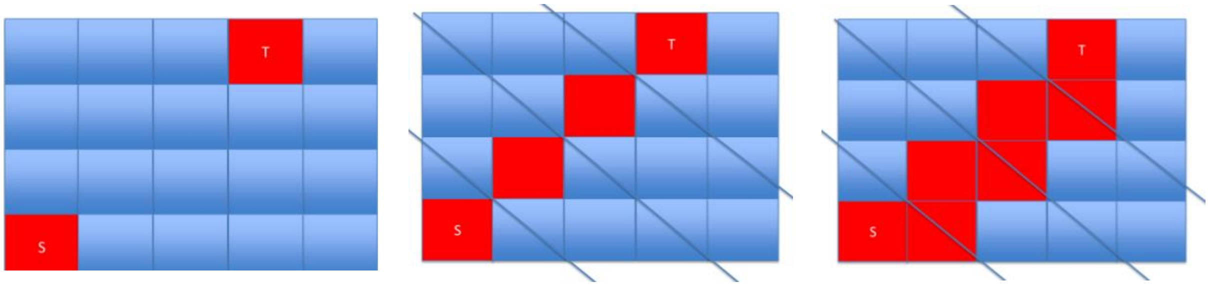
METHODS

Data Acquisition

Images for this study were acquired as part of an IRB-approved prospective adaptive radiotherapy protocol at M.D. Anderson Cancer Center Orlando. The planning kVCTs as well as the repeated weekly kVCTs for all the patients were acquired using a Philips Brilliance CT system (Philips Medical Systems, Best, The Netherlands) or a Siemens Biograph 64 PET/CT system (Siemens AG, Munich, Germany) with 1x1x3 mm^3 reconstructed resolution. Soft tissue and rigid structures were manually contoured in each of the datasets using a commercial contouring tool (MimVista ®Inc., Cleveland, OH).

Structured Volume Generation

The contours represented the structure outlines and were transmitted as an ordered series of contour vertices. The proposed head and neck deformation algorithm required assignments of each voxel to a single structure and so a structured volume algorithm was developed to assign



(a) (b) (c)

Figure 6.1. Schematic description of a modified Bresenham's line algorithm. (a) depicts two contour boundary points S and T. (b) illustrates connecting the boundary points using a Bresenham's line generating algorithm, (c) illustrates connecting the boundary points using the proposed algorithm. The ordinal rays passing through the contour without intersecting the boundary points are illustrated in (b), while the ordinal rays passing through the contour intersecting the boundary points are illustrated in (c).

the voxels based on the contour vertices. The algorithm was as follows:

1. The kVCT data were pre-processed by creating a secondary voxel grid at 5x in-plane resolution of the CT scan, and assigning the secondary grid voxels that overlapped the contour vertices to the contour structures, which is defined as the 3D voxelized volume described by the contour points. These voxels were termed vertex voxels.
2. Voxels that lay between consecutive vertex voxels were assigned to the contour structure using Bresenham's algorithm [18], with the additional constraint that neighboring assigned voxel pairs shared either the same x or y value. Figure 6.1(a) depicts a 2D slice representation with two vertex voxels, referred to as S and T. Figure 6.1(b) depicts the voxel assignments connecting S and T using the original Bresenham's algorithm and Figure 6.1(c) depicts the voxel assignments with the additional constraint. This process was repeated for all vertex voxels, leading to a series of assigned voxels that formed a closed boundary. The assigned voxels were termed boundary voxels.
3. The voxels inside the contour boundary voxels were determined and assigned using a GPU-based algorithm:

   a. The secondary voxel grid was imported to 3D texture memory.

b. For each unassigned voxel, an accumulator value was created and initialized to 0.

c. Parallel rays along the eight cardinal and ordinal directions were passed through each axial plane of the 3D texture.

d. Each ray sampled every voxel it intersected and updated the voxel's accumulator value. When the ray encountered its first boundary voxel, it added 1 or 2 to every subsequent unassigned voxel's accumulator it intersected, for ordinal or cardinal rays, respectively. The process continued until the ray encountered a second boundary voxel (of any contour). At that point, it stopped adding to the accumulator. As the ray met subsequent boundary voxels, the accumulation process repeated until the ray exited the volume.

e. Once all the rays were traversed, the accumulator dataset was analyzed, and voxels with an accumulator value of 11 or 12 (the maximum), were considered part of the contoured structure.

Figures 6.1(b) and 6.1(c) exhibit the ray intersection with the dataset, and illustrate the requirement of the modification to Bresenham's algorithm. While the cardinal rays pass through the boundary voxel initiating the accumulation process, the ordinal rays do not intersect with the boundary and may lead to errors. Figure 6.1(c) depicts the ordinal rays passing through the contoured boundary, illustrating the utility of the structured volume generation algorithm presented above.

4. Finally, the dataset was down-sampled to the original CT resolution, and voxels with greater than 50% of their volume within the contour were considered to be entirely within the contour.

Mass Element Generation

The next step in the biomechanical model generation was to initialize the biomechanical model’s anatomy. We defined the biomechanical model as consisting of a series of connected mass elements with associated mass-spring damping (MSD) connections in a deformation space where it could be deformed and manipulated. Mass elements were generated at the center of each voxel within the structured volume. For an accurate mass element assignment to specific structures we ensured the following two constraints: (a) Each contour set included a body contour that covered the entire head and neck anatomy, ensuring that no mass element inside the body was excluded from being assigned to a contour structure, and (b) No two contour structures overlapped with each other, ensuring that no ambiguity existed in the mass element assignment process. Each of the mass elements were then associated with a Young’s modulus and a Poisson’s ratio based on the anatomy as previously tabulated for parotid glands [19] and other structures [20]. In addition, the damping coefficient for each of the mass elements was set to 0.43 [21, 22].

MSD Connection Initialization Algorithm

Connecting the mass elements with each other using a spring damper formulation ensured that the mass elements could deform in a physically realistic manner. To achieve this, the deformation space was first sub-divided into a 3D uniform cell grid, and each mass element was assigned a hash value (or an identification value) based on the cell that contained it. Mass elements were then sorted by their hash value using a GPU-based fast radix algorithm [23]. Once sorted, a local neighborhood search was performed in a parallelized manner around each mass element (hereafter referred to as the search element in this section), to find nearby mass elements. When a nearby mass element was within a threshold distance (determined by the voxel size of the input CT) from the search element, an MSD connection was established and the nearby mass element became a connected element for the given search element. The steps in this process were as follows: first, an array in the GPU memory was initialized to hold mass element identifiers for the connected elements of each search element. A second array was initialized in the GPU memory to hold rest state orientations for each of the connections, where the rest state orientation was defined as the vector from the search element to the connected element. At the end of this step, the biomechanical model was complete.

Computing Model Deformations

We defined the head and neck deformation to be actuated by user-defined rigid transformations of skeletal structures. While the skeletal structures underwent rigid transformations, the muscles and the soft tissues in the head and neck region underwent elastic deformations governed by internal corrective forces. The internal corrective forces on each mass element were calculated as a summation of tensile spring force, shear spring force, and a dashpot damping force. At rest state, the elastic internal corrective forces were set to 0. When deformed (as further discussed in section 4), the model's mass elements were relocated to new positions inside the deformation space, which caused the internal corrective forces to be non-zero.

The calculation of the internal corrective forces began by computing the tensile spring force, $\vec{f}_{Y,ab}$, between mass elements *a* and *b*[24, 25]:

$$\vec{f}_{Y,b} = \frac{Y_a+Y_b}{2}\left(\frac{|\vec{p}_{ab}|- |\vec{l}_{ab}|}{|\vec{l}_{ab}|}\right)\frac{\vec{p}_{ab}}{|\vec{p}_{ab}|}\ , \tag{1}$$

where Y_a and Y_b were the elastic moduli for mass elements *a* and *b*, respectively, $\vec{l}_{ab}$ was the rest length orientation for MSD connection between mass elements *a* and *b*, and $\vec{p}_{ab}$ was the projection:

$$\vec{p}_{ab} = \frac{\vec{l}_{ab}}{|\vec{l}_{ab}|}\left(\frac{\vec{l}_{ab}\cdot\vec{l}'_{ab}}{|\vec{l}_{ab}|}\right)\ , \tag{2}$$

where $\vec{l}'_{ab}$ was the vector from mass element *a* to mass element *b* in the deformed state.

The shear spring force, $\vec{f}_{S,ab}$, on mass element *a* due to mass element *b* applied along a rejection vector, $\vec{r}_{ab}$, was

$$\vec{f}_{S,b} = -\frac{S_a+S_b}{2}\left(\frac{\vec{r}_{ab}}{|\vec{l}_{ab}|}\right) , \tag{3}$$

where S_a and S_b were the shear moduli for mass elements *a* and *b*, respectively, and

$$\vec{r}_{ab} = \vec{l}'_{ab} - \vec{p}_{ab} . \tag{4}$$

The dashpot damping force, $\vec{f}_{v,ab}$, was calculated from the relative velocities of the mass elements, $\vec{v}_a$ and $\vec{v}_b$, and a local damping factor μ_{ab}:

$$\vec{f}_{v,b} = \mu_{ab}(\vec{v}_b - \vec{v}_a) , \tag{5}$$

The internal corrective force, $\vec{f}_a$, on mass element *a* was then computed by summing over all its spring connections:

$$\vec{f}_a = \textstyle\sum_b\left(\vec{f}_{Y,b} + \vec{f}_{S,b} + \vec{f}_{v,b}\right) . \tag{6}$$

Once the internal forces were computed, the new positions, $\vec{x}_a^{n+1}$, and velocities, $\vec{v}_a^{n+1}$, of the mass elements were updated from the values ($\vec{x}_a^{n}$, $\vec{v}_a^{n}$) at the previous iteration n, using Implicit (Backward) Euler integration[26]

$$\vec{v}_a^{n+1} = \vec{v}_a^{n} + \left(\frac{\vec{f}_a}{m_a} + \vec{g}\right)\delta , \tag{7}$$

$$\vec{x}_a^{n+1} = \vec{x}_a^{n} + \vec{v}_a^{n+1}\delta, \tag{8}$$

where δ was the time step between iterations, m_a was the mass of mass element *a*, and $\vec{g}$ was acceleration due to gravity.

Model Actuation

The next step was actuating the biomechanical head and neck model to represent posture and physiological changes.

Simulating posture changes. Simulating posture changes was conducted using a three-step model actuation procedure. In the first step, we transformed the 3D skeletal anatomy using a graphical user-interface that controlled the individual contoured skeletal structures such as the skull, mandible, and cervical vertebrae. The muscle and the soft tissues were deformed in the second and third steps by applying the soft tissue corrective forces as previously discussed in section 3. During this transformation, the skeletal rotations were constrained at any step to be not more than one degree about a single axis to ensure that the soft tissue deformations occurred in small steps. From a computational perspective, such small soft tissue deformations avoided instabilities that arose from time integration computations [27, 28]. Each of the contoured muscle structures were deformed with the skeletal and other soft-tissue positions as rigid-body constraints. Similarly, the remaining soft tissue deformations outside any contoured structures were computed with the muscle deformation and skeletal transformations being taken as rigid-body constraints. The deformation process was repeated until all the soft tissue structures reached equilibrium deformations.

Simulating physiological changes. Physiological changes were incorporated on a 3D structured volume basis using a two-step iterative approach. For illustration purposes, we present a scenario where a planning target volume (PTV) underwent regression with all the other structure volumes undergoing normal elastic deformations. In the first step, the PTV regression was computed with its surrounding structures providing a rigid-body constraint. For a given change in the PTV volume, the surface area of the PTV volume was reduced in a physically accurate manner by decreasing the rest length of each connection inside the PTV. It was performed as follows: Before a volume change was initiated inside the PTV, its mass elements were clustered to form a set of cuboids. The volume of the set of cuboids was summed to represent the PTV volume.

Using the initial cuboid volume V_i and the volume change v_i for a cuboid i, the change in rest length a_i was computed using

$$(l + a_i) * (w + a_i) * (h + a_i) = V_i + v_i, \tag{9}$$

where l, w, and h were the length width and height of each cuboid. From a physiological perspective, the rest length change a_i demonstrated the loss or gain in volume. In order to maintain a stable deformation, the rest length change was constrained to be not greater than 2 mm per iteration. The change in the rest length led to internal elastic corrective forces that subsequently deformed the PTV to reflect the volume regression.

In the second step, the reduced PTV was considered as a rigid body constraint and the remaining soft tissue anatomy surrounding the PTV was deformed as previously explained in section 3. The two-step iterative process continued until the entire anatomy deformation converged.

Generating kVCT Images Representing the Deformed Anatomy

Due to the deformed anatomy, there was not a one-to-one correspondence between mass elements and image voxels. This led to the following issues: (a) gap artifacts where voxels enclosed no mass elements but had a transiting MSD connections, (b) hole artifacts, and (c) aliasing artifacts arising from skeletal head and neck rotations. These three issues were addressed by using the following steps:

1. A new data volume was initialized that represented the synthetic CT image of the biomechanical model in the deformed state. The data volume dimensions and resolutions were set to be the same as that of the reference kVCT image.
2. For every mass element in the biomechanical model, the current position was converted from the deformation space into a voxel address in the new data volume.
3. Each of the new voxels was assigned the Hounsfield intensity (HU) originally associated with the enclosed mass element, or the average HU of multiple enclosed elements.

4. To address the hole/gap artifacts,
    a. Ray-traces were defined between every mass element along MSD connections, with interval sampling equal to half of the MSD connection's rest length.
    b. For every mid-MSD sampled position in the deformation space, the corresponding position in the new data volume was evaluated to see if it was in an empty voxel.
    c. When the mid-MSD sampled position was in an empty voxel, the HU value was interpolated from the two mass elements connected by the MSD connection with the interpolation weighted by their relative distances from the sample position.
    d. If one of the mass elements was skeletal anatomy, its corresponding weight was set to 0 to ensure the new data volume maintained the same rigid skeletal structure shape as that of the reference kVCT.
5. To reduce the aliasing artifacts that occurred during head rotation, we employed a GPU-based linear intensity smoothing technique[29] . The method worked as follows:
    a. The new data volume holding the synthetic CT anatomy was loaded into the GPU's texture memory, and sampled at twice its current resolution.
    b. The tri-linear intensity smoothing was then applied on the up-sampled data to remove the aliasing artifacts. The intrinsic interpolation removed the artifacts from the feature edges.
    c. This interpolated image was then down-sampled to the original dimensions to produce the final synthetic CT.

RESULTS

We now present our results on the model development from a reference head and neck kVCT anatomy and its associated contoured structures. The structure volume generation algorithm precisely associated the head and neck voxels to their corresponding contoured structures.

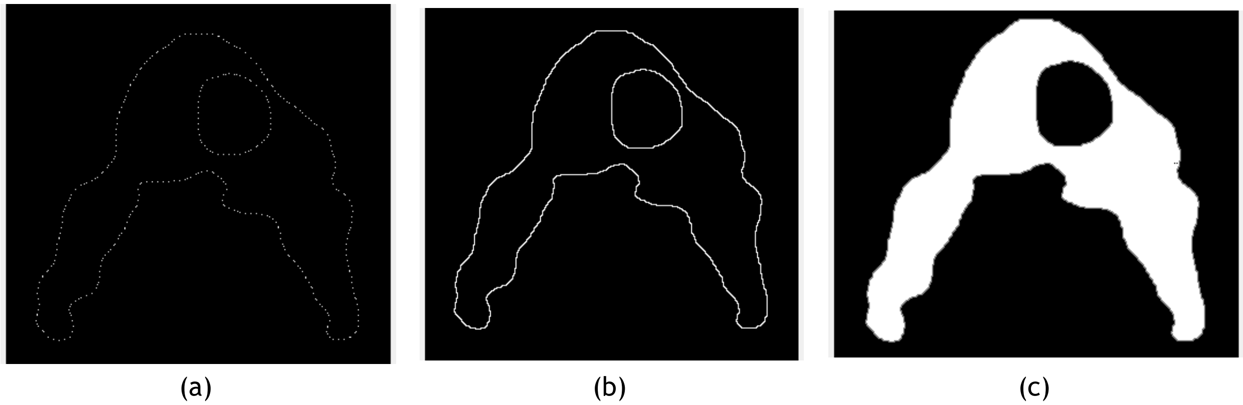


Figure 6.2. Contour filling algorithm. (a) A slice of contour points obtained from a DICOM RT structure. The contour lines that connect all the contour points are shown in (b). (c) The volume filled contour structure.

Figure 6.2(a) shows a planar set of disconnected contour points associated with internal and external boundaries. Figure 6.2(b) shows the connected contour boundary generated from the planar contour points. The algorithm was able to form nested and distinct boundaries. Figure 6.2(c) shows the final structure volume with voxels associated with the structure represented in white colored pixels. It can also be seen that the structure volume generation algorithm accurately distinguished the structures inside the inner-most circular boundary to not belong to the contoured structure volume and not associate them with the structure.

Figure 6.3(a-c) illustrate the head and neck model developed from the reference kVCT. Figure

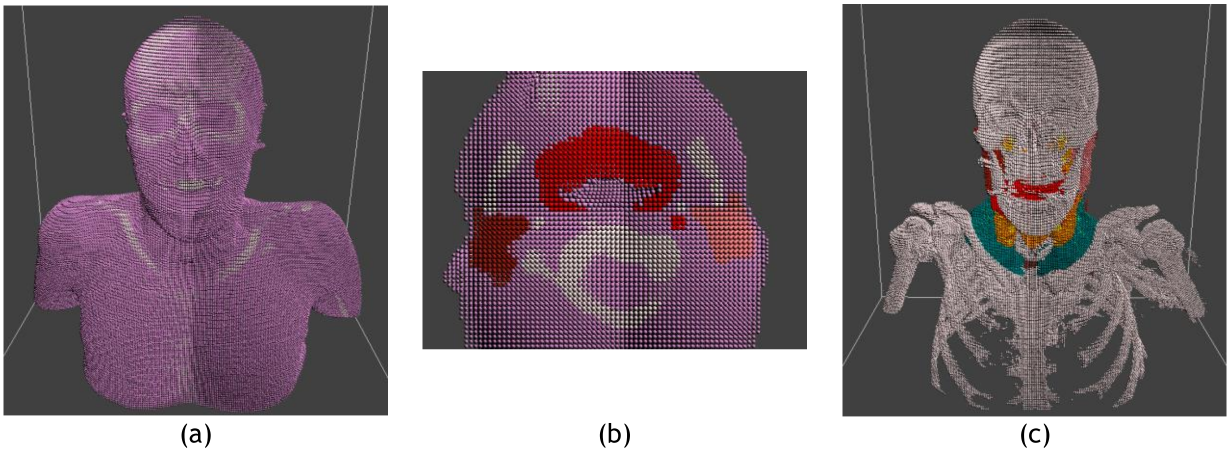
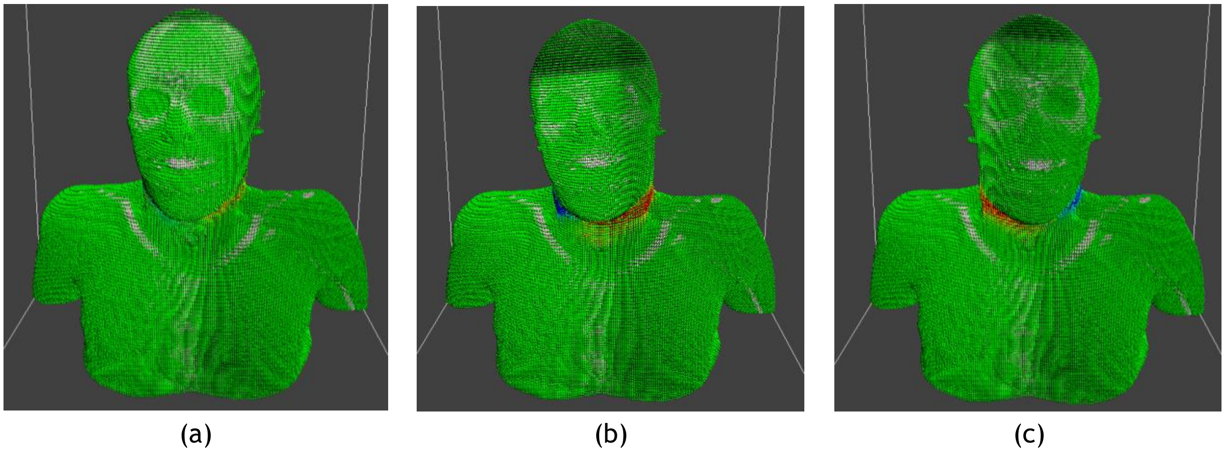


Figure 6.3. The model in its rest position is shown with the entire anatomy (a), and the critical contours (c). A 2D slice of the neck region showing the PTV (in red), the surrounding soft tissue (violet) and the rigid bone structure is illustrated in (b).

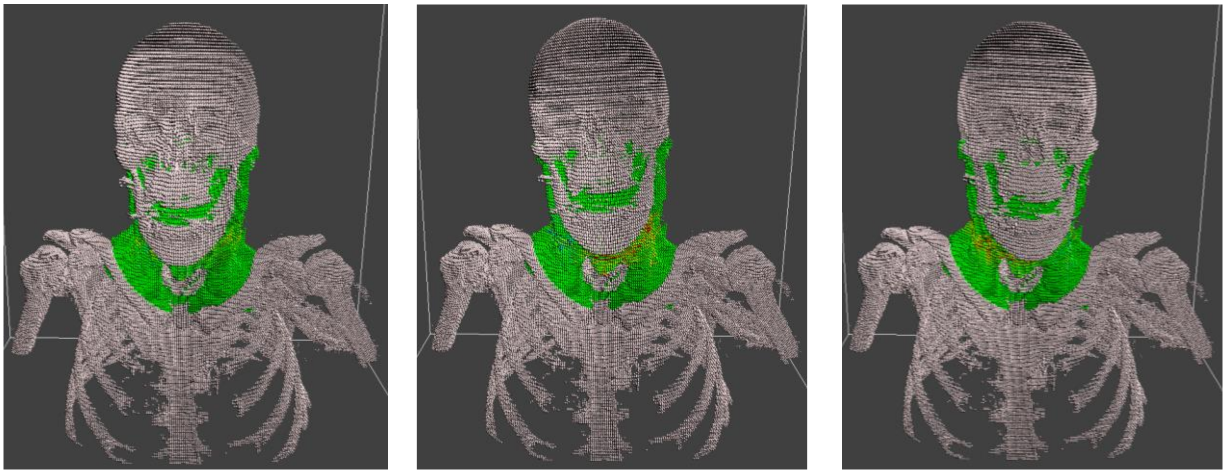


(a) (b) (c)

Figure 6.4. Biomechanical deformation with all the anatomical structures in the head and neck region. The model before the deformation is shown in (a). Two different neck rotations are demonstrated in (b) and (c).

6.3(a) shows the rest state of the biomechanical model. The rigid skeletal anatomy is shown in white mass elements while the deformable soft tissues are shown in violet mass elements. The volumetric nature of the biomechanical model is shown in figure 6.3(b) using a 2D slice of the 3D deformable model anatomy that included the contour filled PTV elements (red) and the bone anatomy (white). The parotid glands on either side of the PTV are also shown in red and pink colored mass elements. The structure volume generation also enables selecting structures undergoing deformation and coupling them with the bone anatomy. Figure 6.3(c) shows the critical contoured structures along with the skeletal anatomy. Each of the contoured structures is shown using a distinct color representation.

Figure 6.4(a-c) shows the biomechanical deformation caused by skull and neck discs rotation along the caudal-cranial axis. The local strain observed in the anatomy was color-coded to represent soft tissue contraction (green-blue) and stretch (yellow-red) as quantitated in table 6.1. The soft tissue regions that underwent neither a contraction nor a stretch were colored green. Deformation differences showed the subtle soft tissue and muscle deformations caused by changes in the patient posture during radiotherapy treatment. They also showed the model's fidelity in representing the 3D deformations that a head and neck anatomy undergoes during different postures. Nevertheless, a visual evaluation of the deformations observed for



(a) (b) (c)

Figure 6.5. Biomechanical deformation with only critical radiotherapy structures in the head and neck region. The model before the deformation is shown in (a). Two different head and neck rotations are demonstrated in (b) and (c).

each of the postures also suggested the qualitative accuracy of the observed deformations. Three examples of the deformation associated with head and neck posture changes are presented in figure 6.5. Figure 6.5(a-c) shows the deformation of a biomechanical model consisting of the head and neck skeletal structure and a PTV, parotid glands and neck muscles (figure 6.3(c)). Other muscle structures were excluded for this simulation. The biomechanical deformation of the critical structures, in this case, was caused by skull and cervical vertebrae rotation along the body axis. The strains associated with each of the mass elements are color coded as previously discussed in table 6.1. The strains signified a local stretch and contraction that occurred during such skull rotations representing posture changes. The differences in the deformation also showed the muscle deformations caused by changes in the patient posture.

Table 6.1 Normalized color-coding scheme used for representing the strain. R, G, and B represent the red, green, and blue color channels.

Displacement (mm)	R	G	B
-9.9	0.0	0.0	1.0
-6.6	0.0	0.5	1.0
-3.3	0.0	1.0	1.0
0	0.0	1.0	0.0
3.3	1.0	1.0	0.0
6.6	1.0	0.5	0.0
9.9	1.0	0.0	0.0

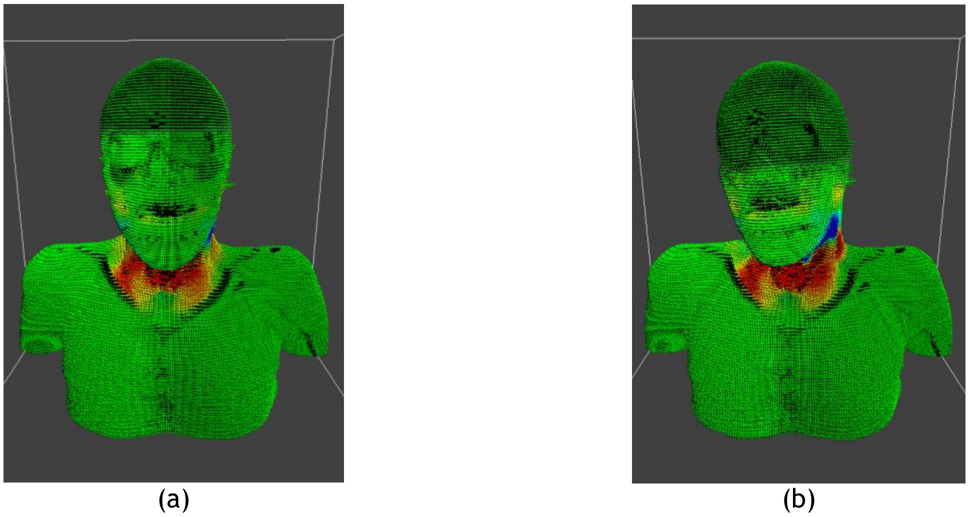
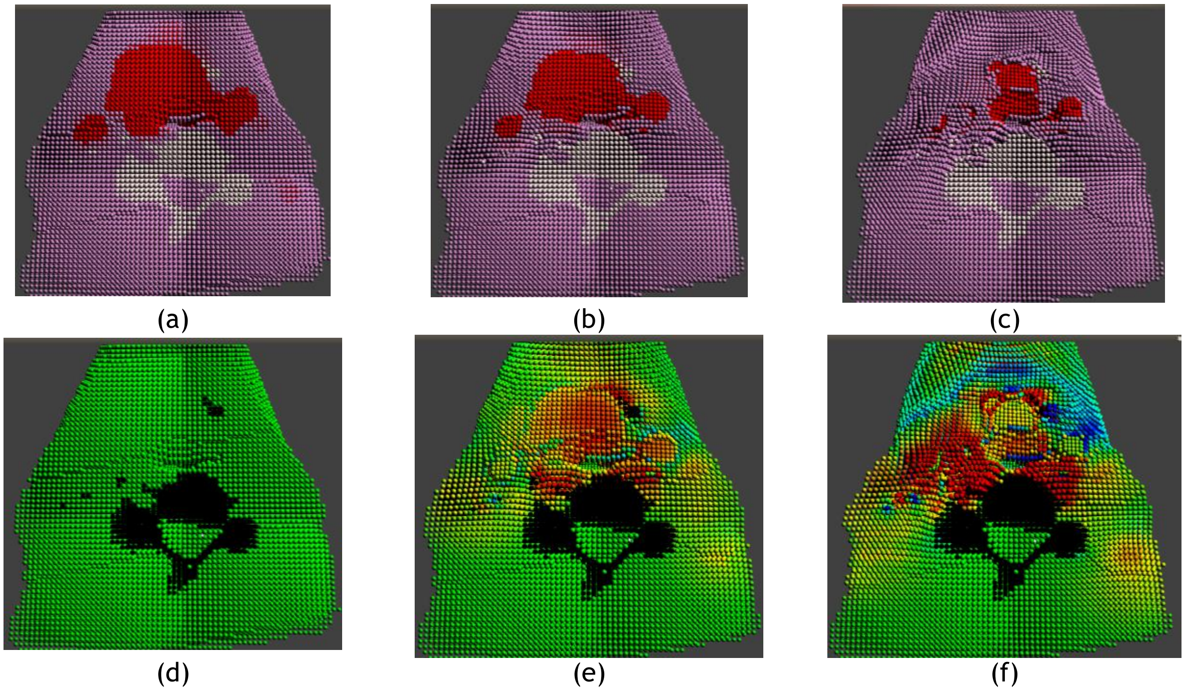


Figure 6.6. Biomechanical deformation with strain depicted by a color-coded heat map. The model before the posture change is shown in (a). A different head and neck posture is demonstrated in (b).

We now present the model results for simulating physiological regression using the reference biomechanical model (shown in figure 6.3(a)). The PTV was shrunk by 30% to simulate tumor regression. The local distribution of the regression is shown in figure 6.6(a), illustrating the local variations in the tissue expansion and contraction. Figure 6.6(b) shows the head and neck deformation when the head and neck skeletal structure was rotated by angle of 10 degrees, both local tissue stretching (red-orange color) and tissue compression (blue color) can be seen during head and neck rotation with simulated PTV regression.

Figure 6.7(a-c) shows a slice of the biomechanical model in the neck region with the PTV in red. Figure 6.7(d-f) show the corresponding stress in the muscle when compared to the original anatomy state. Figure 6.7(b) shows the internal model changes caused by simulating a PTV shrinkage of 10%. The corresponding deformation strain is shown in figure 6.7(e). Similarly, figure 6.7(c) shows the internal model changes caused by simulating a PTV shrinkage of 30% with the corresponding deformation strain shown in figure 6.7(f). A significant amount of deformation outside the PTV can be observed for each of the cases signifying the role of biomechanical head and neck model. The local distribution of the regression demonstrated the local variations in the tissue expansion and contraction for



(a) (b) (c)

(d) (e) (f)

Figure 6.7. A 2D cross section of the model illustrating tumor regression. Rest state (a) and deformed state representing 10% (b) and 40% (c) PTV volume reduction are shown. The corresponding color-coded strain maps for the three states are shown in (d, e, and f).

known PTV regression.

Figure 6.8 illustrates how the artifacts are compensated for during generation of the simulated kVCT images. The source CT used to generate the model is shown in figure 6.8(a). After a 45 degree rotation, the hole artifacts are abundant in figure 6.8(b). Figure 6.8(c) shows an image after ray-tracing is used to fill the holes, but jagged edges are still apparent. The texture based smoothing algorithm is applied in figure 6.8(d) to produce the final simulated kVCT data set.

Simulated kVCT images corresponding to different states of deformation and rotation are as shown in Fig 6.9a-d. Specifically, figure 6.9(a) shows the 2D slice generated from the model at rest state. The corresponding slice with 30% PTV regression is as shown in figure 6.9(c). The underlying non-rigid deformation of the anatomy during the neck rotation is evident in figure 6.9(b) with no PTV regression and figure 6.9(d) with 30% PTV regression. The differences in the deformations stemmed from changes in the PTV region. Such deformations can only be

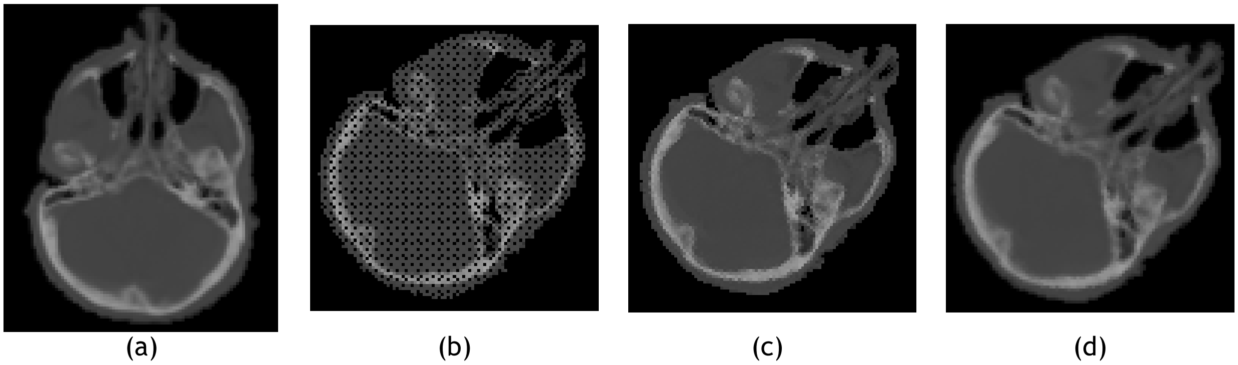


Figure 6.8. Simulated kVCT generation and artifact correction. An axial slice of the source kVCT used to generate the model is displayed in (a). After rotating the head by 45°, the resultant model generated image is full of holes and aliasing as shown in (b). Holes are addressed by raytracing between model elements, and filling holes with an interpolated intensity (c). Aliasing is addressed using a texture based smoothing algorithm to produce the final image(d).

simulated using a high-resolution physics-based deformation model, which is a key contribution of the paper.

Model Validation

In order to validate the numerical accuracy of the biomechanical model, first the general deformation mechanics were tested by comparing the soft tissue response to local global loads with analytic ground truth calculations. Next, the head and neck models were validated by inducing clinical posture changes in the bony anatomy and comparing the soft tissue deformations induced in the model with the deformations recorded in the clinical data. Lastly, the effect of tumor regression was validated by applying a clinically observed deformation to the tumor and comparing the soft tissue deformations induced in the model with the clinical data.

Soft Tissue Response to Local and Global Loads

The local load simulation examined the model deformations by applying a known amount of force using a 5 cm radius rigid sphere onto a 10x10x10 cm^3 cubic piece of tissue and determining the subsequent deformation using the model. The ground truth deformation was

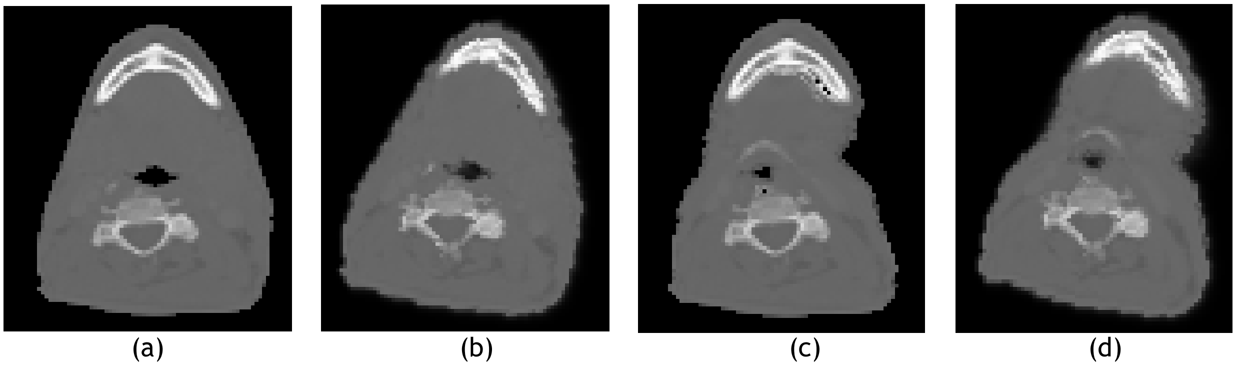


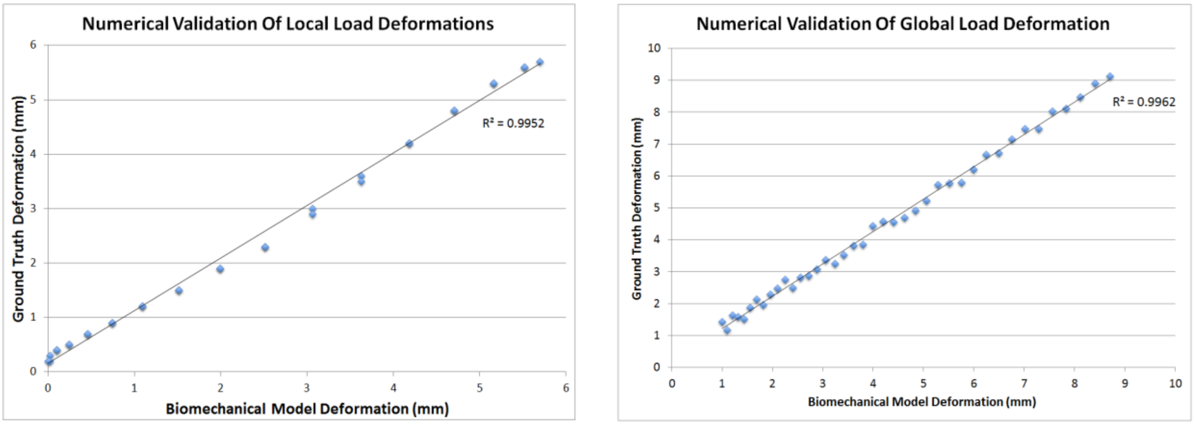
Figure 6.9. Simulated kVCT slice at the tumor target. Rest state (a) and rotated by 10 degrees (b) are shown. The kVCT slice representing the same anatomy with the PTV reduced by 30% is shown in (c). The deformation of the model where the skull is rotated by 10 degrees is shown in (d).

computed using a classical Euler beam theory, which is a simplified method of calculating deflection due to a load using the linear theory of elasticity [30]. The two deformations were compared along the row of soft tissue voxels that lay on the cube surface and intersected the contact point between the sphere and the cube. The deformation computed using the proposed system matched the ground truth with a R^2 fit value of > 0.98, as shown in figure 6.10(a).

The global load simulation dealt with scenarios where the entire soft tissue structure was applied with an uniform load (e.g. gravity)[30]. In the first of two scenarios, the top layer of the soft tissue was anchored and a gravitational force was applied. The resulting deformation was compared with the numerical solution provided by Barber[30]. In the second scenario, the direction of gravity was reversed, which led to tissue compression. The model predicted deformation matched well with the ground truth with an R^2 > 0.98 for both scenarios, the results of the hanging mass are shown in figure 6.10(b).

Reproducing Clinical Deformations

For the purpose of producing ground truth data to be used for DIR validation in head and neck radiotherapy, the biomechanical model needed to be able to reliably reproduce the type of deformations typically seen in the clinic. To validate the model's ability to do this, 10 patients



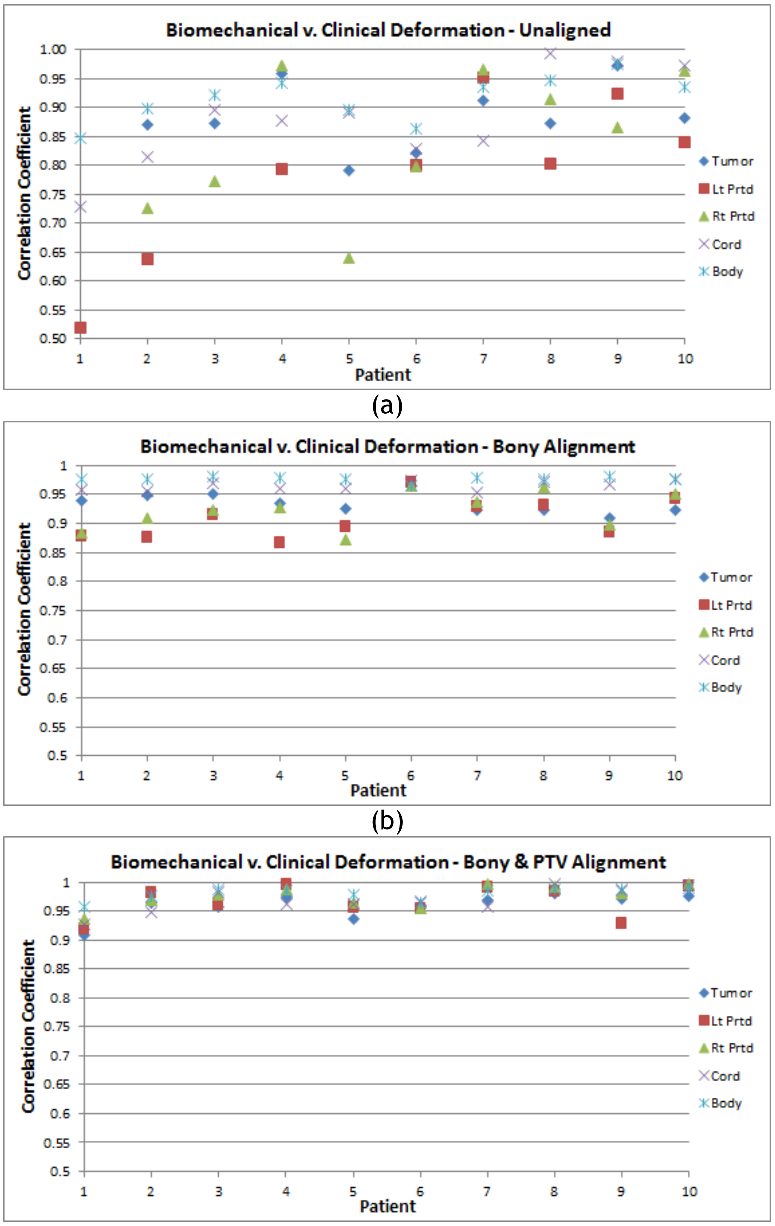
(a) Local Load Simulation (b) Compressed Mass - Global Load Simulation

Figure 6.10. Elastostatic validation study using local and global force application. A numerical comparison of the elastostatic displacement for a column of voxels is plotted against ground truth computed using a Green's function solution. (a) shows the response of a cube of soft tissue to a local load applied in the form of a spherical mass. (b) shows the response of a cube of soft tissue in response to being compressed by gravity.

were selected that had weekly kVCT scans over the course of their radiotherapy treatment. An image registration was performed between the initial planning kVCT and a kVCT acquired during the final week of the patient's treatment to obtain the changes in the skeletal positions in the head and neck region. A biomechanical model was assembled from the initial planning CT, and the deformation vectors obtained from the DIR for the skeletal anatomy were applied to the model's skeletal anatomy. This forced the model into the posture of the final week's kVCT, while allowing the soft tissue to deform in response to the changes in skeletal anatomy. An image-based analysis was then performed comparing the planning CT, the final weekly kVCT, and the model-generated kVCT equivalent. The Pearson product-moment correlation coefficient [31], was used as the image metric for this analysis. Equation (10) shows the correlation coefficient, with X_i and Y_i representing a voxel intensity of the ground-truth (final week's kVCT) and test data (model-generated kVCT) sets, while $\bar{X}$ and $\bar{Y}$ represent the mean intensity of the corresponding data sets.

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \qquad (10)$$

To establish the baseline correlation, the planning kVCT was first correlated with the final



(a)

(b)

(c)

Figure 6.11. Correlation of model generated data sets with induced soft tissue deformation from posture changes with clinically observed deformation. (a) shows the baseline correlation between the planning CT used to generate the biomechanical model and the target CT used as the deformation endpoint. (b) shows the correlation after applying posture changes to the model to match the weekly CT. (c) shows the correlation after the inclusion of tumor regression to the model.

weekly kVCT. Analysis was performed on each contoured structure, as well as the entire head and neck region. The results for the primary tumor target, left and right parotids, the spinal cord, and the body are shown in figure 6.11(a) for each of the ten patients. The correlation varies greatly from patient to patient, as the standard deviation for the average correlation

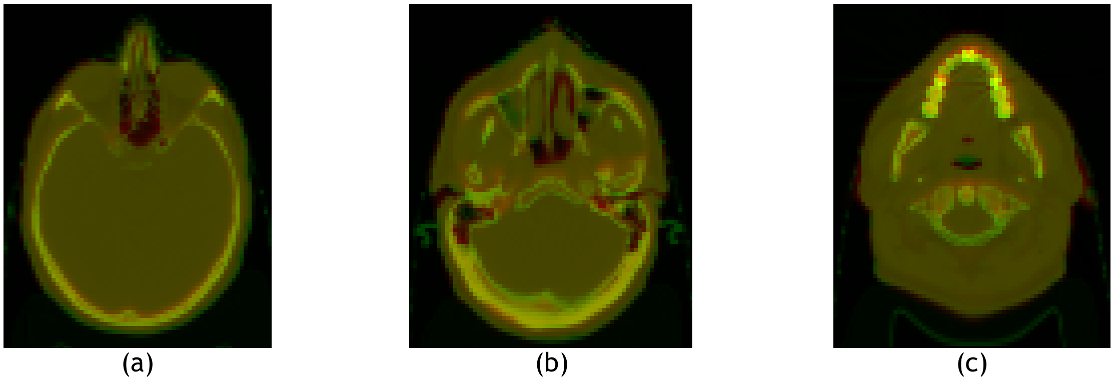


Figure 6.12. Model generated kVCT (red) images overlaid on weekly kVCT (green) images at three axial levels. Yellow areas show good agreement between the model generated images and the weekly CT images. Areas tinted more green or red show a disagreement where one image had a higher intensity.

of a structure approached 20%. For patient 1, the tumor and right parotid had correlations below 0.5.

The correlation between the model generated data set and the final weekly CT is shown in figure 6.11(b). The displacement of soft tissue voxels ranged from 9.3 to 24.3 mm, with an average maximum of 15.3 ± 4.9 mm. The correlation increased significantly in all cases. The average correlation of the tumor increased from 0.844 ± 0.136 to 0.934 ± 0.017.

This experiment, however, doesn't allow for physiological changes such as tumor regression. The tumor targets for these patients reduced in volume by an average of 5.1 ± 3.6 mL, with the largest regression in patient 5 at 9.26 mL. The volume changes also induced a shift in the tumor center of mass. The maximum shift was in patient 1 at 15.9 mm, while the average displacement of the center of mass in all 10 patients was 4.44 mm. To validate the model's response to tumor regression, the deformation vectors from the DIR were also applied to the primary tumor target as well as the posture changes to the skeletal anatomy. The correlation between the model generated data with tumor regression and the weekly CT are shown in figure 6.11(c). The average correlation coefficient increased from the previous case where only posture changes were introduced. The correlation of the primary tumor target increased to 0.960 ± 0.022.

The model was able to very closely reproduce the soft tissue anatomy seen in a spectrum of clinical patients when posture changes were induced, as seen in figure 6.12. When tumor regression was also set to match the clinical data, the correlation increased even further, such that the lowest correlation of the structures analyzed was still greater than 0.9. This experiment illustrated that the model is capable of simulating physics-based deformations that are very close to clinically seen deformations, and validated the model's ability to ultimately generate ground-truth deformations to be used for DIR validation studies.

DISCUSSION

DIR plays a pivotal role in the head and neck adaptive radiotherapy but validation of various DIR algorithms has been hampered by the lack of a quantitative high resolution ground truth. In this paper, we presented a GPU based high-resolution biomechanical head and neck model using kVCT images that can be used to overcome this difficulty. The biomechanical model will be used for generating CT equivalent 3D volumes that simulate posture changes and physiological regression in order to validate image-guided patient positioning approaches, for example DIR accuracy of different registration paradigms. The model can also be generated using other volumetric imaging modalities, such as megavoltage CT (MVCT) and cone beam CT (CBCT). Moreover, with the advent of Magnetic Resonance Imaging for both patient simulation and on-board imaging, the model will be an effective tool for generating deformed MRI images and verifying registration accuracy.

It has been previously established that head and neck anatomy involves a complex musculoskeletal feedback system. In this paper, we specifically focus on deforming the muscle and soft tissue system based on known skeletal positions and orientations. The model actuation effectively simulates head and neck deformation from patient posture and physiological regression. The model is currently actuated using a simple graphical user interface that individually controls the skeletal structures in the head and neck region. Such a framework enables us to model soft tissue and muscle deformations representing posture changes. There are multiple methods for introducing anisotropic volume regression. Two simple implementations for the model presented in this paper are adding multiple contours inside a single anatomy and individually controlling the volume regression of the sub-structures, and heterogeneously manipulating the elastic properties of the spring connections of an element. To the best of our knowledge, this is the first biomechanical model capable of simulating head and neck physiological changes such as volume regression.

The biomechanical head and neck model discussed in this paper employed a mass-spring approach to deform the head and neck anatomy. While several CPU based mass-spring modeling approaches exist, the model discussed in this paper is unique as it employs a very efficient GPU based approach to deform the head and neck anatomy (discussed in section II.D and II.E).

One of the salient features of the biomechanical model is its real-time ability to deform. Using state-of-art graphics processing units, it was observed that the model was able to deform at a rate of 60 deformations per second. While the real-time nature of the model may not have a direct impact on the DIR validation, we envision that it will have a significant impact for on-line adaptive radiotherapy where DIR plays a key role. Recent advancements in image segmentation, registration and online adaptive planning has led to systems that can perform their tasks in real-time. Thus having a biomechanical model guided validation that can match the speed provided by these algorithms will also be essential for future developments in adaptive radiotherapy.

Future work will focus on improving the skeletal model to simulate more physically and physiologically realistic articulation. The biomechanical properties in our model were obtained from the literature, but these properties vary between patients. We will develop a technique for estimating patient specific tissue elastic properties by inversing the forward deformation model for known deformations. This will provide patient specific head and neck biomechanical models which will be useful for adaptive radiotherapy. While others have used low resolution finite element models to estimate elastic properties, the proposed high resolution model with its complex musculoskeletal behavior will provide a more accurate estimation. The GPU based platform presented in this paper enables the complex calculations to be performed in parallel and in a scalable fashion in nearly real-time.

REFERENCES

[1] Institute, N. C. *NCI Fact sheet on head and neck cancer*. Available from: http://www.cancer.gov

[2] Bentzen, S. M., Constine, L. S., Deasy, J. O., Eisbruch, A., Jackson, A., Marks, L. B., Ten Haken, R. K., and Yorke, E. D., "Quantitative analyses of normal tissue effects in the clinic (QUANTEC): An introduction to the scientific issues.," International Journal of Radiation Oncology Biology Physics, S3-S9 (2010).

[3] Britton, K., Dong, L., and Mohan, R., "Image Guidance to Account for Interfractional and Intrafractioin Variations: From a Clinical and Physics Perspective," in [Image-Guided Radiotherapy of Lung Cancer], J. Cox, J. Chang, and R. Komaki, Editors, Informa Healthcare USA, Inc.: New York, NY (2007).

[4] Li, X. A., Liu, F., Tai, A., and Ahunbay, E., "Development of an online adaptive solution to account for inter- and intra-fractional variations," Radiotherapy and Oncology 100(3), 370-374 (2011).

[5] Liu, Erickson, Peng, and Li, "Characterization and Management of Interfractional Anatomic Changes for Pancreatic Cancer Radiotherapy," International Journal of Radiation Oncology, Biology, Physics 83(3), e423-e429 (2012).

[6] Stewart, J., Lim, K., Kelly, V., and Xie, J., "Automated Weekly Replanning for Intensity-

Modulated Radiotherapy of Cervix Cancer," International Journal of Radiation Oncology, Biology, Physics 78(2), 350-358 (2010).

[7] Bujold, A., Craig, T., Jaffray, D., and Dawson, L., "Image-Guided Radiotherapy: Has It Influenced Patient Outcomes?," Seminars in Radiation Oncology 22(1), 50-61 (2012).

[8] Wang, J., Bai, S., and Chen, N., "The clinical feasibility and effect of online cone beam computer tomography guided intensity modulated radiotherapy for nasophyrngeal cancer," Radiotherapy Oncology 90, 221-227 (2009).

[9] Sifakis, E., Neverov, I., and Fedkiw, R., "Automatic determination of facial muscle activations from sparse motion capture marker data," ACM Transactions on Graphics 24(3), 417-425 (2005).

[10] Dilorenzo, P. C., Zordan, V. B., and Sanders, B. L., "Laughing out loud: Control for modeling anatomically inspired laughter using audio," ACM Transactions on Graphics 27(5), 125:1-8 (2008).

[11] Albrecht, I., Haber, J., and Seidel, H., "Construction and animation of anatomically based human hand models," ACM Siggraph, 98-109 (2003).

[12] Tsang, W., Singh, K., and Fiume, E., "Helping hand: An anatomically accurate inverse dynamics solution for unconstrained hand motion.," ACM Siggraph, 319-328 (2005).

[13] Santhanam, A., Willoughby, T., Shah, A., Meeks, S. L., Rolland, J. P., and Kupelian, P., "Real-time Simulation of 4D lung tumor radiotherapy using a breathing model," Lecture Notes on Computer Science 11, 710-717 (2008).

[14] Komura, T., Shinagawa, Y., and Kunii, T. L., "Creating and retargeting motion by the musculoskeletal human body model," The visual computer 16(5), 254-270 (2000).

[15] Veress, A. I., Segars, W. P., Weiss, J. A., Tsui, B. M., and Gullberg, G. T., "Normal and pathological NCAT phantom data based on physiological realistic left ventricle finite element models," IEEE Transactions on Medical Imaging 25(12), 1604-1616 (2006).

[16] Lee, S. H. and Terzopoulos, D., "Heads Up! Biomechanical modeling and neuromuscular control of the neck," ACM Transactions on Graphics 25(3), 1188-1198 (2006).

[17] Harders, M., Hutter, R., Rutz, A., Niedere, P., and Szekely, G., "Comparing a simplified FEM approach with the mass-spring model for surgery simulation," Studies in Health Technology Informatics 94, 103-109 (2003).

[18] Bresenham, J. E., "Algorithm for Computer Control of a Digital Plotter," IBM Systems Journal 4(1), 25-30 (1965).

[19] Arda K, Ciledag N, Aktas E, Aribas B, and K, K., "Quantitative assessment of normal soft-tissue elasticity using shear-wave ultrasound elastography," American Journal of Roentgenology 197, 532-536 (2011).

[20] Huang L., Bakker N., Kim J., Marston J., Tis J., and Cullinane D., "A multi-scale finite element model of bruising of soft connective tissue," Journal of forensic biomechanics 3 (2012).

[21] Shi, H., *Finite Element Modeling of Soft Tissue Deformation*, in *Department of Electrical and Computer Engineering*. 2007, University of Louisville: Louisville, Kentucky.

[22] Tilleman, T., Tilleman, M., and Neumann, M., "The elastic properties of cancerous skin: Poisson's ratio and Young's modulus," Israel Medical Association Journal 6, 753-755 (2004).

[23] Harris, M., Sengupta, S., and Owens, J. D., "Parallel prefix sums (scan) with cuda," in [GPU Gems 3] (2007).

[24] Boresi, A., [Elasticity In Engineering Mechanics], 3 ed, (2010).

[25] Love, A. E. H., [A Treatise On Mathematical Theory Of Elasticity]: Dover Publications, (2011).

[26] Hwu, W.-m., [GPU Computing Gems Jade Edition], 1 ed, Vol, 2, (2011).

[27] Rougier, E., Munjiza, A., and John, N. W. M., "Numerical comparison of some explicit time

integration schemes used in DEM, FEM/DEM and molecular dynamics," International Journal for Numerical Methods in Engineering 61, 856-879 (2004).

[28] Mesit, J., Guha, R., and Chaudhry, S., "3D Soft Body Simulation Using Mass-spring System with Internal Pressure Force and Simplified Implicit Integration," Journal of Computers 2(8), 34-43 (2007).

[29] Sigg, C. and Hadwiger, M., "Fast Third-Order Texture Filtering," in [GPU Gems 2], M. Pharr and R. Fernando, Editors, Addison-Wesley Professional (2005).

[30] Barber, J. R., ed. *Elasticity 3rd edition*. Solid mechanics and its applications. 2009, Kluwer academic publishers.

[31] Rodgers, J. L., Nicewander, J. L., and Nicewander, W. A., "Thirteen Ways to Look at the Correlation Coefficient," American Statistician (42), 59-66 (1995).

CHAPTER 7: GPU-based Modeling and Simulation of Interactive Patient-Specific Hyper-Elastic Head and Neck Deformations

Abstract

Patient specific biomechanical models have many potential applications in domains ranging from medical simulations to animations. To be used effectively, they must be fast, accurate, and robust. In this paper, we present a patient-specific head-and-neck biomechanical model with hyper-elastic material properties that satisfies the fast, accurate and robustness criteria for medical applications. The high resolution patient geometry was instantiated from clinical patient imaging. Results show that the biomechanical deformations were achieved at interactive frame rates. The soft tissue deformation response was realistic for large posture changes involving dramatic rotations of the head, and substantial volume changes such as severe weight loss or tumor regression. The model was integrated with an in-house volume renderer, producing visualizations of the deformation using the original intensity information from the patient imaging, moving towards clinical incorporation of such models.

INTRODUCTION

Biomechanical models have progressed at a remarkable rate in recent years, allowing physically accurate deformations and visually impressive rendering. Models that describe large posture changes and major actuations such as running and walking have contributed to realism in the gaming industry. Based on the model resolution and the complex geometry, the speed, robustness, and accuracy varies from one simulation to the other. Our focus is on the medical applications related to cancer treatment, where the greatest importance is given to (a) the accuracy of the soft tissue deformation response, (b) the real-time nature of the computations, and (c) the usage of patient-specific geometry. There are many potential applications for such models, and have been investigated for application domains such as the virtual surgery simulation. Physics-based methods, such as finite element and mass-spring, allow for a broad array of simulations, from gross posture changes to subtle day-to-day deformations like tumor regression. Mass-spring systems typically employ a linear elastic material model, providing fast, stable deformations. However, biological tissues exhibit a hyper-elastic response beyond small deformations [1, 2]. Finite element models can provide physiologically realistic deformations, but they can be computationally expensive and typically apply a tetrahedral meshing which may lower model resolution.

Biomechanical Models in Radiation Therapy

In the field of radiation therapy, biomechanical models have not yet found a place in daily clinical activities. Major contributors hindering their incorporation are specificity to the patient, volumetric resolution, and model's computational complexity.

Patient specificity. Treatment planning and quality assurance in radiation therapy depends heavily on repeated patient imaging, such as computed tomography (CT) or magnetic resonance (MR). Every patient's anatomy differs from one to another and so the biomechanical model needs to be instantiated on a subject specific basis. To be clinically relevant, the model geometry must be instantiated from such imaging modalities. The complexity in creating patient specific models lies in the variability of anatomy between patients when attempting to fill the internal meshing structure. Imaging in radiation therapy is typically accompanied by a set of contoured structures to help delineate the anatomy, but the model must still be geometrically robust to establish a consistent internal structure.

Model geometry resolution. Instantiating model geometry directly from CT or MR images creates high-resolution, patient-specific biomechanical models. The model resolution needs to account for model geometry resolution at 1-3 mm. For even a small anatomy such as the head and neck region, the number of elements tend to be closer to 1 million with steep gradient changes to the model's biomechanical properties.

Computational complexity. The usage of a high resolution model geometry coupled with a hyper-elastic constitutive model renders the head and neck deformations as a computationally complex task. While the usage of graphics processing units (GPUs) offers scope for addressing this computational task, the methodology has not been previously investigated.

Patient-specific modeling of biomechanical deformations will be clinically irrelevant without interactive manipulation and accurate deformations. To address the above-mentioned limitations with intent on facilitating biomechanical modeling into the radiotherapy clinic, we developed a mechanism to couple high-resolution patient-specific geometry with a hyper-elastic finite element material model, to allow clinically realistic deformations at interactive frame rates. The key contributions of this paper are (a) the GPU-based implementation of a hyper-elastic material model, and (b) the usage of high resolution head-and-neck geometry obtained from clinically acquired and contoured CT images, approaching one million elements, while maintaining interactive framerates. To our knowledge, such a high resolution biomechanical model of head-and-neck has not previously been investigated.

Related Work

Biomechanical models of human anatomy have been developed for applications ranging from computer animation [3] to CT image registration [4]. The high complexity of the human head-neck-trunk musculoskeletal system is caused by the highly constrained spatial relationship among the large number of articular bones (57) and muscle actuators. Such models have been used to model complex human motion in the face [5-7], the neck [8], the torso [9, 10], the hand [11-14], and the leg [15, 16]. Biomechanical model development of head and neck anatomy [3, 7] have been developed for physically-realistic (qualitative) animation applications which involve motion ranges that are significantly greater than those found in radiotherapy treatment setup variations. Other biomechanical models of the lungs [17-19] have been used for modeling the radiation dose delivered to lung tumors and surrounding tissues during the treatment.

The computational complexity of biomechanical human anatomy models has resulted in mostly non-real-time performance. For cases such as the lungs, where the breathing motion can be pre-computed to some extent, real-time performance using a linear elastic model has been demonstrated [20]. To date, biomechanical hyper-elastic models, specifically of head-and-neck anatomy, that can deform in real-time and incorporate high resolution geometry have not been developed.

We hypothesize that biomechanical modeling approaches can be made quantitative and encompass the relatively limited range of motion found in radiotherapy treatment setup variations, using template models and 3D CT scans as inputs to assemble the patient-specific model. As a first step in this direction, we investigate the development of a GPU-based hyper-elastic biomechanical model for the head-and-neck anatomy.

METHODS

Hyper-Elastic Constitutive Model

Most biological tissues exhibit a hyper-elastic response, i.e., they are virtually incompressible but able to undergo large elastic deformations. Maintaining an interactive framerate for a hyper-elastic material model implementation is much more difficult than a linear Hookean model [21].

Hyper-elasticity is formulated by deriving a strain-energy function from the deformation gradient tensor, which is defined as the partial derivative of the deformed state with respect to the reference state. For implementation, a generalized Ogden material model was chosen, which defines the strain energy, *W*, in terms of the principal stretches, λ_i, and the shear modulus, μ [22].

$$W = \sum_{p=1}^{N} \frac{\mu_p}{\alpha_p}\left(\lambda_1^{\alpha_p} + \lambda_2^{\alpha_p} + \lambda_3^{\alpha_p} - 3\right) \quad (1)$$

$$2\mu = \sum_{p=1}^{N} \alpha_p \mu_p \quad (2)$$

The Ogden model allowed variations with a variety of strain-energy functions by adjusting the parameters *N* and α, such as Neo-Hookean (*N* = 1, α = 2)[23] and Mooney-Rivlin (*N* = 2, α_1 = 2, α_2 = -2)[24, 25].

The principal Cauchy stresses, σ_i, can be found from the 2nd Piola-Kirchoff stress tensor, $\boldsymbol{\tau}$, which is derived from the partial derivative of the strain energy function with respect to the principal stretches[26].

$$\sigma_i = \lambda_i \tau_i = 2\lambda_i \frac{\partial W}{\partial \lambda_i} = \sum_{p=1}^{N} \mu_p \lambda_i^{\alpha_p} \quad (3)$$

Hyper-Elastic Implementation

The structure established by the proposed model's meshing algorithm allows direct calculation of the principal stretches, λ_i, as the rest state vectors are stored as components along the principle axes.

The deformation of the local volume around each element, *a*, was found by examining its connected nearest neighbors, *b*, and comparing the current state ($\vec{l}'_{ab}$) with the rest state orientation ($\vec{l}_{ab}$). The deformation vector for each element was deconstructed into the projection ($\vec{p}_{ab}$) along the rest state vector and the corresponding rejection ($\vec{r}_{ab}$).

$$\vec{p}_{ab} = \frac{\vec{l}_{ab}}{|\vec{l}_{ab}|}\left(\frac{\vec{l}_{ab} \cdot \vec{l}'_{ab}}{|\vec{l}_{ab}|}\right) \qquad (4)$$

$$\vec{r}_{ab} = \vec{l}'_{ab} - \vec{p}_{ab} \qquad (5)$$

The principle stretch was then defined as the sum of the squares of the differences between the current state and the rest state, while maintaining the directionality by normalizing the projection and rejection components.

$$\vec{\lambda}_{ab}^{\alpha_p} = \left[\left(\frac{|\vec{p}_{ab}| - |\vec{l}_{ab}|}{|\vec{l}_{ab}|}\right)\frac{\vec{p}_{ab}}{|\vec{p}_{ab}|}\right]^{\alpha_p} + \left[\frac{\vec{r}_{ab}}{|\vec{l}_{ab}|}\right]^{\alpha_p} \qquad (6)$$

The force on element *a* was then calculated by summing over the contributions of its connected elements *b* according to equation (3), where *A* is the cross-sectional area of the interaction between elements.

$$\vec{f}_a = A \sum_b \vec{\sigma}_{ab} = A \sum_b \sum_{p=1}^{N} \mu_p \vec{\lambda}_{ab}^{\alpha_p} \qquad (7)$$

A GPU implementation of the hyper-elastic constitutive model was achieved by assigning each element a separate thread. Each thread then looped over all connections for its assigned element, accumulating the force as described by equation 7. Improved stability was found by establishing an array to hold the reciprocal force on each element *b* due to element a, $\vec{f}_{b|a}$, as equal and opposite to the force calculated on element *a* due to element *b*. The reciprocal forces were then summed in a subsequent GPU kernel. The total internal corrective force applied to element *a* was then equal halves the directly calculated force and the total accumulated reciprocal force.

Integration Scheme

From the principal Cauchy stress at each element, the internal force vectors, $\vec{f}_a$, can be computed [27], and the new positions, $\vec{x}_a^{n+1}$, and velocities, $\vec{v}_a^{n+1}$, of the mass elements updated from the values ($\vec{x}_a^n$, $\vec{v}_a^n$) at the previous iteration n, using Implicit (Backward) Euler integration. To improve robustness and stability, at a compromise with performance, the trapezium rule was applied to the implicit integration scheme according to Heun's method[28]. Integration was implemented on the GPU by consolidating data using the zip iterators of the Thurst library, and feeding into a custom developed functor operator.

Model Instantiation and Meshing

The input for the geometry consisted of patient-specific CT images of the head-and-neck anatomy with each of the sub- structures (e.g. tumor, glands, muscles) contoured by a clinician. This data acquisition and contouring is performed as part of regular patient treatment. The patient imaging was initially resampled to be isotropic. Contoured structures were then loaded and run through a volume filling algorithm on the GPU involving multi-directional ray-casting in order to tag voxels by their respective parent structures [29]. All elements were localized in the deformation space on a one-to-one

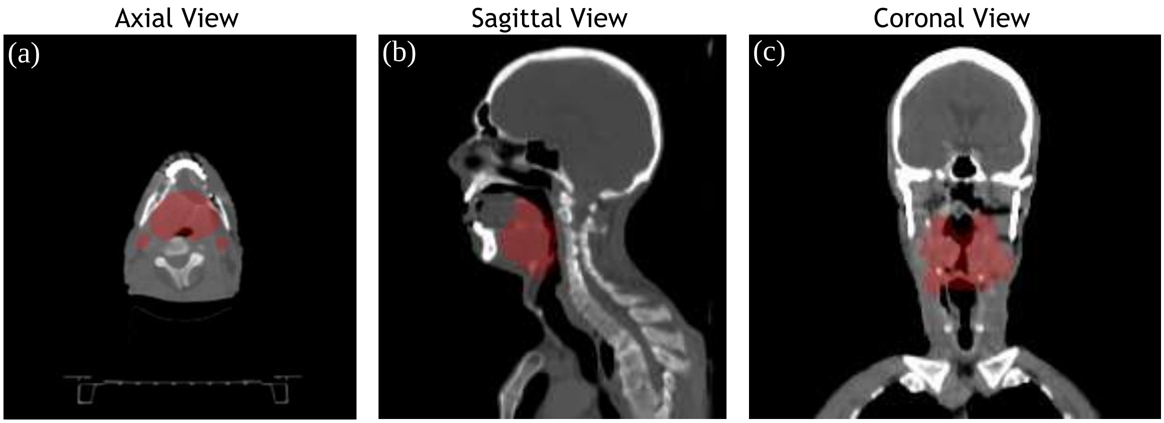


Figure 7.1. Patient CT with contoured structure of tumor target overlaid in red on the (a) axial, (b) sagittal, and (c) coronal views.

correspondence with the imaging voxels, but sub-systems of elements could then be controlled independently based on their assigned parent structures.

The meshing algorithm was optimized to run on GPU, to facilitate on-demand adaptive re-meshing between deformation iterations to accommodate prolonged deformations. A uniform sub-division of the deformation space was established with cells on the same order as the imaging voxels. Elements were first assigned hash values according to their residing cell. Elements were reorganized using the sorting operator of the Thrust library, to optimize memory access patterns on the GPU. Once sorted, a thread was launched per element to perform a local neighborhood search. Up to 26 connections were established isotropically about each element in a 3x3x3 cube, prioritizing nearest neighbors and applying limiting criteria.

Actuation / User Control

The model was able to be controlled by a series of keyboard and mouse controls. Actuating the head-and-neck involved rotating the skeletal anatomy, cranium and cervical vertebrae, and allowing the soft tissues to deform according to their connections with the skeletal anatomy and the internal corrective forces. The user chose an axis and used keyboard keys to rotate the head in one degree increments. The connection vectors describing the rest state configuration were also transformed to reflect the new posture.

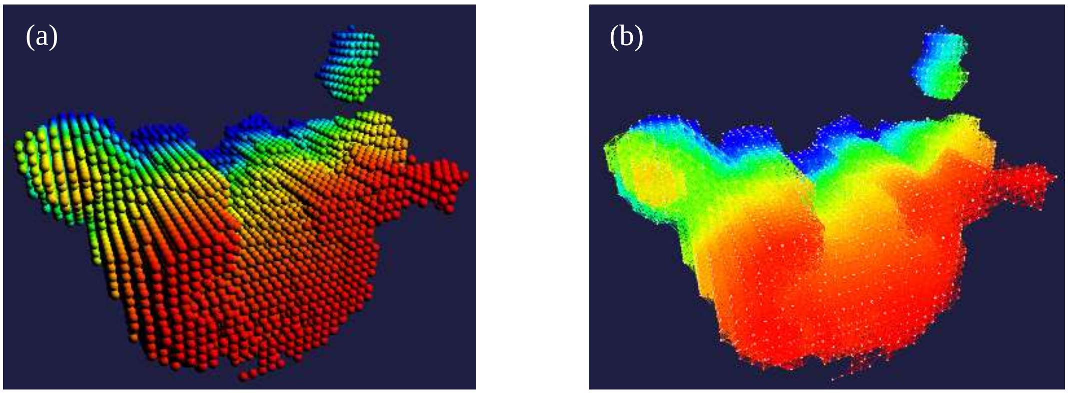


Figure 7.2. Illustration of the particle system (a) and results of the meshing algorithm (b) for contoured structure of tumor target.

Volumetric changes were applied by changing the rest length magnitude of connections between elements in the rest state. Although the entire model was established as a single meshed system, specific structures could be manipulated independently. To cause regression in specific organ, such as the tumor target, the user cycled through available contoured structures to choose the tumor target, then manipulated a slider, adjusting a multiplier which was applied to the rest state magnitude. It should be noted then, that a 50% reduction to the rest state connections for a cubic element would result in an 87.5% reduction in volume. Similarly, general weight loss was simulated by reducing the rest lengths of all the connections for the general soft tissues.

Implementation Environment

All development was performed in Ubuntu 12.04 LTS environment using C/C++. GPU implementation utilized the CUDA 6.5 toolkit, and made extensive use the Thrust libraries. Direct rendering of the biomechanical system was done using openGL, while the volume rendering was developed using gtkmm libraries.

RESULTS

The results and renderings presented in this section were created from a single patient CT scan. Table 7.1 describes the CT data set and the composition of the biomechanical model created from it after everything outside the external body contour had been removed.

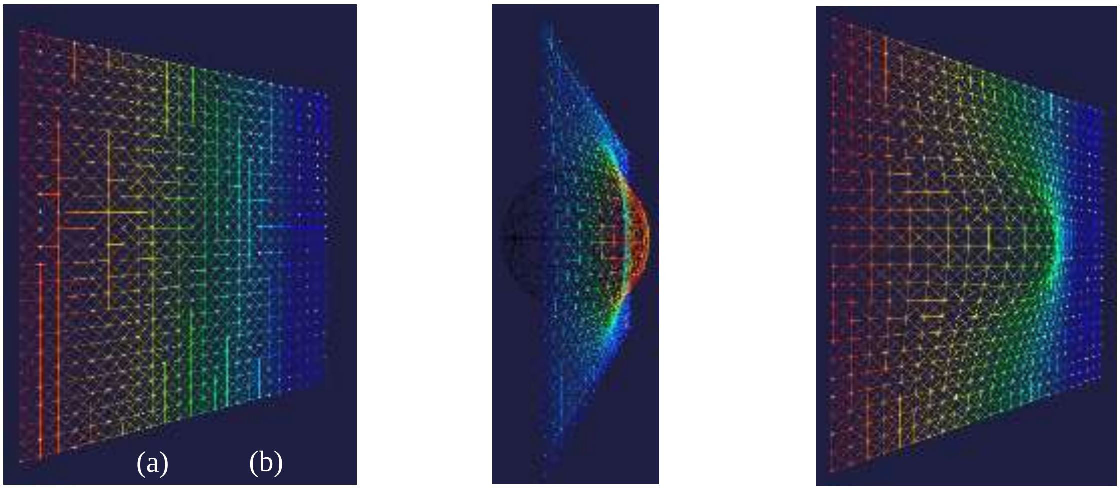


Figure 7.3. Illustration of how re-meshing incorporates previous deformations into the model's rest state configuration (a), eliminating and strain/force contributions from the prior deformations (b), and creating a new rest state configuration (c).

Figure 7.2 demonstrates the robustness of the meshing algorithm with respect to irregular input geometry. Here the tumor target structure has multiple distinct parts, and many concavities. The structure contains 12,990 elements, displayed in figure 7.2(a), with 292,736 interconnections between them, displayed in figure 7.2(b). The meshing algorithm for a model this size computed in 20-25 ms on average.

Table 7.1. Model composition. Instantiated from CT imaging with 216^3 voxel with 2.5 mm isotropic resolution.

	Elements	Connections
Entire Model	835,082	17,749,495
Skeletal	149,351	1,054,834
Soft Tissues	685,731	16,742,661

Figure 7.3 illustrates the applications for adaptive re-meshing. A system of elements was instantiated as a two dimensional sheet, as shown in figure 7.3(a). The sheet was then deformed by a user-controlled spherical object, producing the force map displayed in figure 7.3(b). While in the deformed state, the system was re-meshed, incorporating the deformation into the system's rest state configuration, such that the internal corrective forces returned to zero. Thus, when the spherical object was removed, the sheet remained at rest in the deformed state, bulging in the center, as seen in figure 7.3(c). Thus, for large deformations, that involve a significant change in the mesh topology, a frequent re-meshing ensures the deformation smoothness.

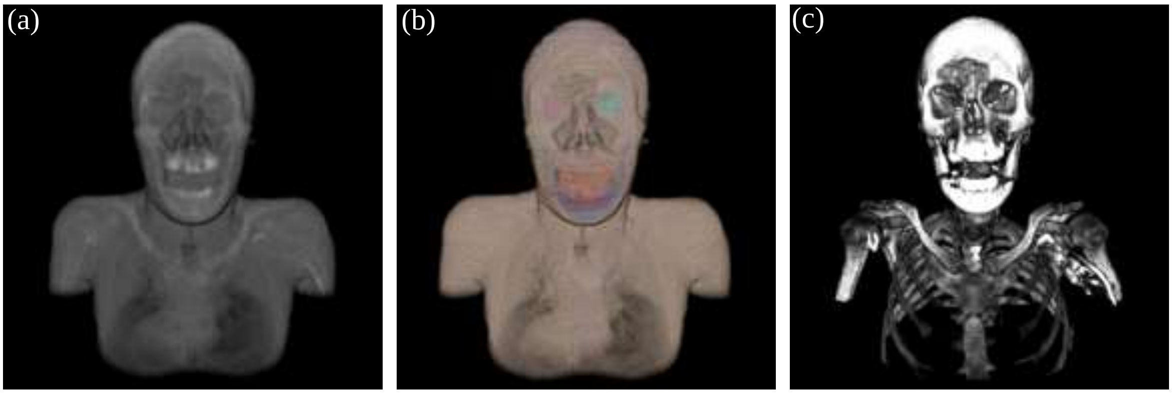


Figure 7.4: Patient specific biomechanical model displayed as CT equivalent intensity (a), contoured structures (b), and skeletal system (c).

The meshing algorithm was tested systematically to measure the performance with respect to the number of elements and total connections established. It was found that the total computation time for the meshing was equivalent to 2 µs per element, or 80 ns per connection. These two values are also related directly by the average number of connections per element in the model of just under 26. Meshing for the complete model described in table 1 completed in just over 1350 ms on average.

The complete model, consisting of over 800,000 elements with nearly 18 million connections between them, deformed at over 18 frames per second (fps) using a Nvidia GTX 780 Ti graphics card. Each frame in this case represent the entire model deformation for a pre-determined time step. The data from the biomechanical model was then transferred by socket to a volume renderer as point cloud data containing the positions and colors for each element. The additional overhead of transferring the data, localizing the point cloud and updating the buffers for the volume renderer’s raytracing algorithm lowered the frame rate to 12.5 fps.

Figure 7.4 displays the output of the volume renderer, displaying a semi-transparent rendering with original CT intensity, the anatomy delineated according to selected contoured structures, and a slightly more opaque skeletal anatomy found by windowing and levelling the output according to the CT intensity.

Figure 7.5 illustrates clinical scenarios where adaptive re-meshing could be applicable. Here the model was instantiated with only the skeletal anatomy and the tumor target

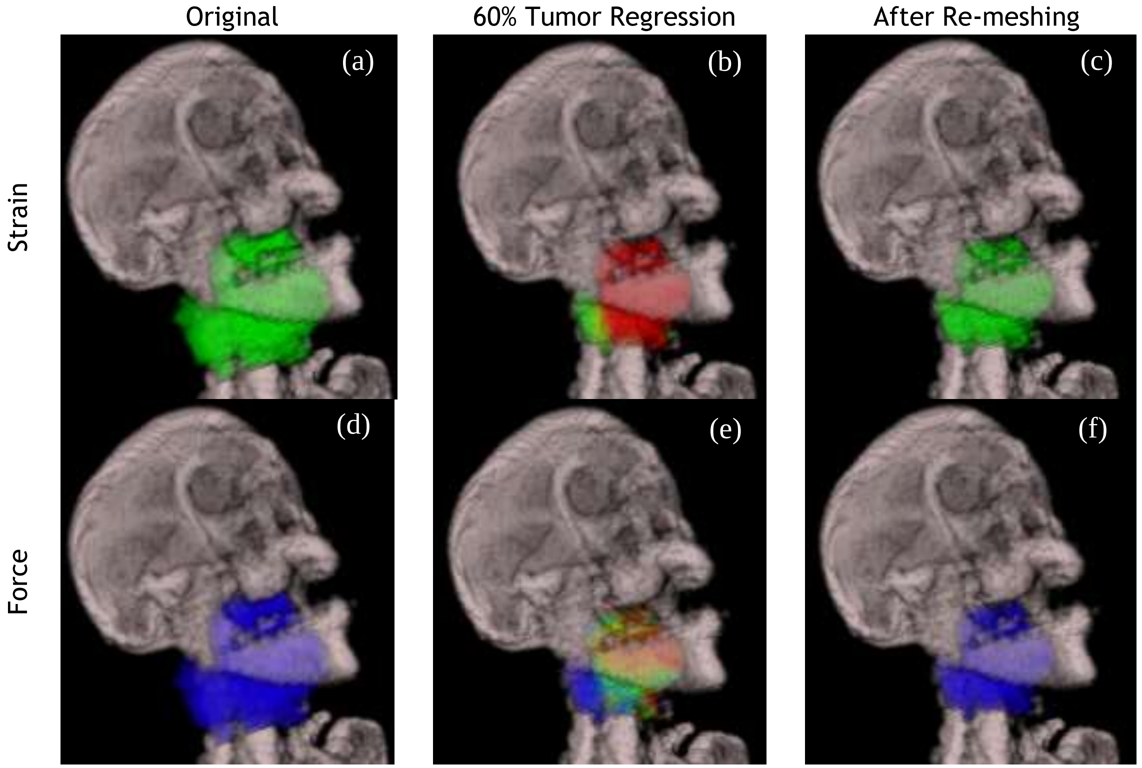


Figure 7.5: Differences in force distribution between the linear elastic material model and the proposed hyper-elastic implementation for a variety of posture and volume changes. The color map in (a-c) indicates areas at rest in green, with compression depicted towards red and elongation towards blue. For (d-f), a heat map was applied to the normalized force magnitude.

structure. After applying a large regression of approximately 60% to the tumor, significant strain and force can be seen acting within the soft tissues, as seen in figures 7.5(b) and 7.5(e), respectively. Before applying further deformations, such as posture changes, the strain and force due to regression can be eliminated by re-meshing, creating a new rest state configuration, as shown in figures 7.5(c) and 7.5(f). Therefore, the deformation due to posture changes will not be influenced by the previous volumetric changes within the model.

Figure 7.6 shows the results after large volume changes were applied to the complete biomechanical model. The first column displays the original anatomy. The second column shows a regression of nearly 90% to the primary tumor, causing contraction in the surrounding soft tissues of the neck. The third column displays general weight loss of approximately 40% applied to the general soft tissues, compressing them over the skeletal anatomy.

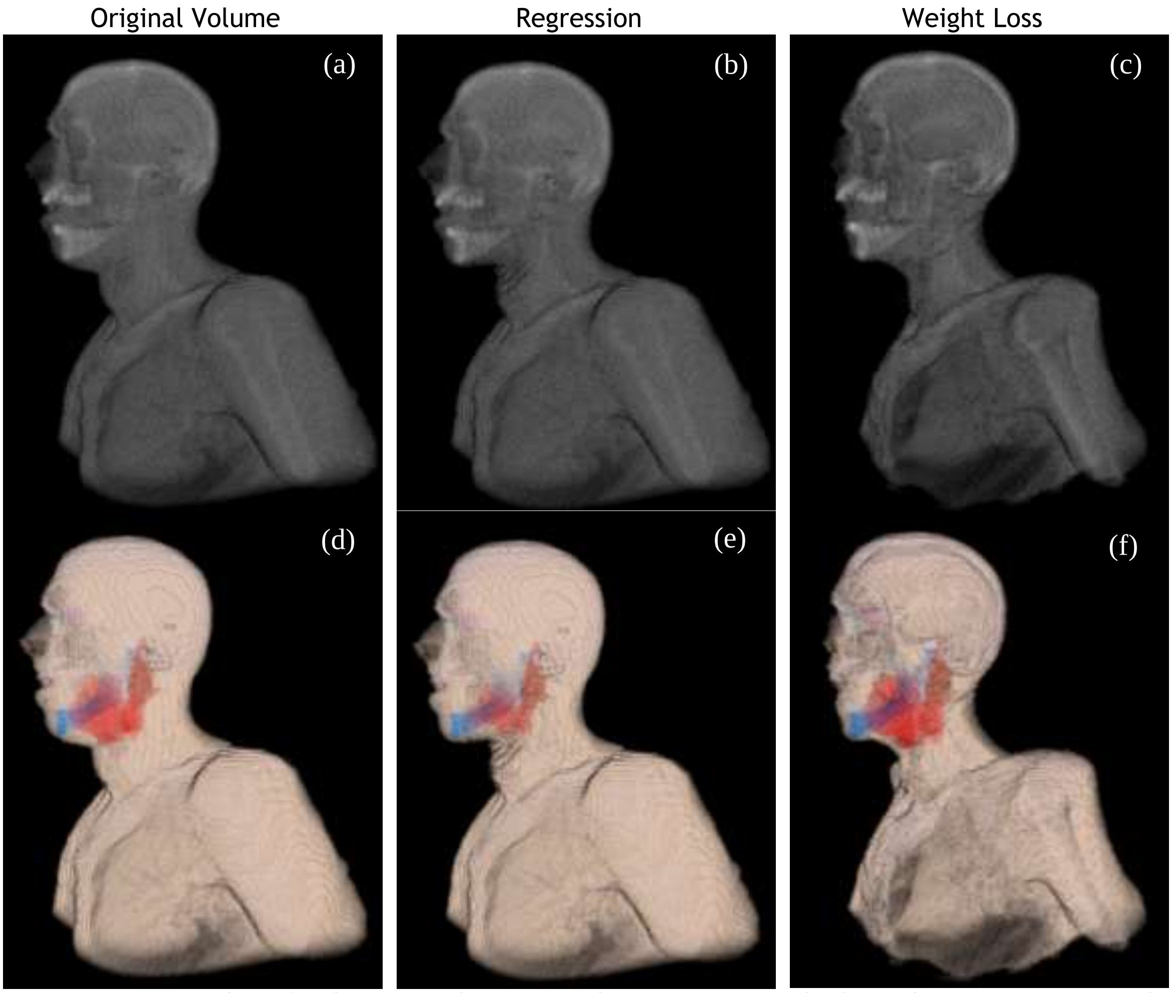


Figure 7.6. Examples of volumetric changes to the anatomy. (a,d) show the CT intensity and contoured structure renderings for the original anatomy. (b,e) show regression of the primary tumor target. (c,f) show severe weight loss applied to the generic soft tissues.

Figure 7.7 builds on the results of figure 7.2 by further incorporating posture changes to the head. The user was able to apply rotations about each axis in one degree increments. In the figure, the rows correspond to the volume changes applied in figure 7.2, with the original anatomy in row 1, 50% regression of the primary tumor in row 2, and 15% general weight loss in row 3. Columns 2 and 3 apply a posture change to each of the volumetrically altered anatomies. Column 2 applied a 25-degree rotation to the left about the cranial-caudal axis. Column 3 applied 15 degree rotations about each axis, tilting the head to the right, bending forward at the neck, and turning to the right.

Figure 7.8 illustrates the differences between the linear elastic material model and the proposed hyper-elastic material model implementation. The figure displays a heat map of the magnitude for the internal corrective forces of the soft tissues, after experiencing a

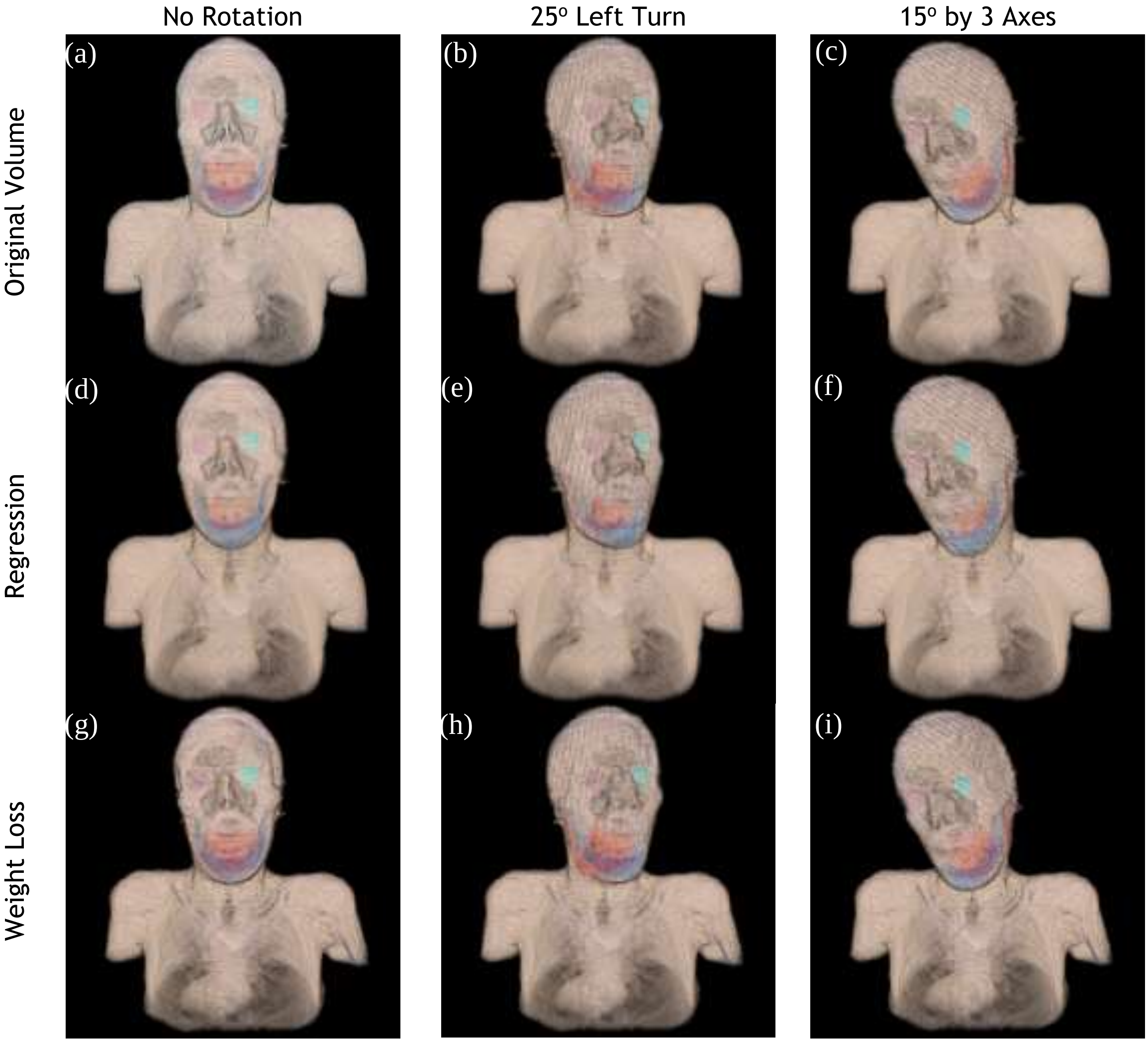


Figure 7.7. Renderings of the contoured structure model for a variety of posture and volume change combinations.

variety of posture changes and volumetric regression. The hyper-elastic implementation produced significantly larger forces in areas of higher strain for all three scenarios presented. The differences observed illustrate the necessity for a hyper-elastic material model when dealing with larger deformations where a linear approximation would no longer be adequate.

DISCUSSION and CONCLUSION

A biomechanical model of the head and neck anatomy was presented in this paper, employing hyper-elastic soft tissue response while maintaining an interactive framerate and responsive rendering. This work focused mainly on the soft tissue response for given

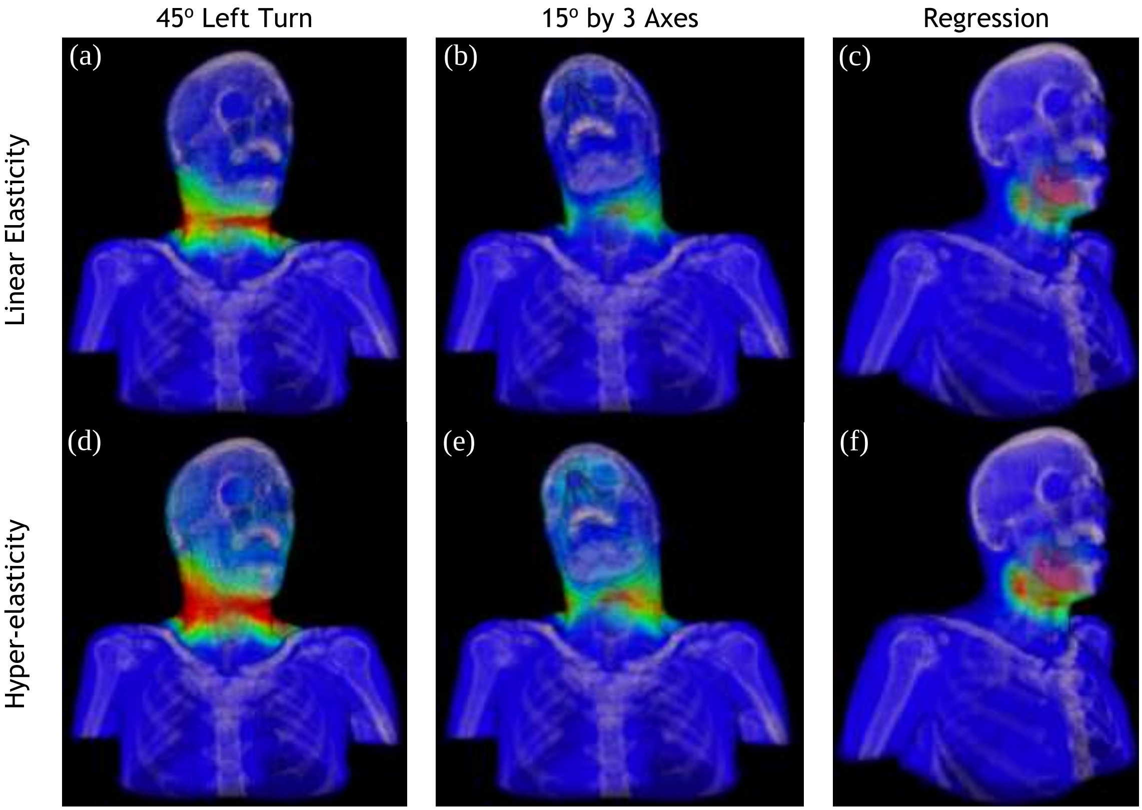


Figure 7.8: Differences in force distribution between the linear elastic material model and the proposed hyper-elastic implementation for a variety of posture and volume changes. A heat map was applied to the normalized force magnitude.

skeletal actuations, and obtaining a fast, accurate, and robust hyper-elastic implementation. Initial results suggest that the model is capable of simulating the posture changes and volumetric variations observed in the patients clinically from day to day over the treatment course.

The model discussed in this paper employs a hyper-elastic constitutive model for representing the soft tissue deformations. While such a model can more closely represent the actual tissue deformations, computing the hyper-elastic properties will be critical in achieving the correct simulation. Future work will focus on model-guided estimations of these parameters. Several such estimations have been investigated in the field of elastography for linear elastic deformations. To date, the computational complexity of hyper elastic deformations has limited a detailed patient-specific estimation of the hyper elastic material properties. Using the model implementation discussed in this paper, we can enable such hyper-elasticity estimations to be performed within realistic computing time.

The computation time observed for the head and neck model was 18 frames per second. While this computational speed clearly satisfies clinical modeling and simulation requirements, improvements may be preferred for tasks such as model-guided deformable image registrations and virtual reality based visualization so. Thus future work will focus on distributing the deformation tasks to multiple GPUs thereby improving the frame rates furthermore.

Future work will also focus on establishing a more efficient pipeline between the biomechanical model and volume renderer. Specifically, we will focus on customizing the user interface, including actuation controls and motion constraints as the models application broadens beyond head-and-neck anatomy. For instance, clinical on-board X-Ray imaging systems can acquire the 3D

skeletal description in real-time and be directly used to apply patient motion to the model and predict internal soft tissue position.

Additionally, the incorporation of camera-based in-room monitoring frameworks to track patient motion and automate model actuation will be explored. This would allow the biomechanical model to be deformed for usage in inter- and intra-fractional motion monitoring and management, providing an estimation of internal changes to the anatomy and physiology without requiring additional volumetric imaging of the patient, avoiding additional radiation exposure.

From an animation perspective, the head-and-neck biomechanical modeling and simulation can be coupled with several applications that require a realistic simulation of human body movements. For instance, animations requiring weight-loss or increase in body mass can be simulated in a physically realistic manner.

With these improvements and further developments, patient-specific, interactive biomechanical models move another step closer to clinical implementation.

REFERENCES

[1] Fung, Y. C., "Elasticity of soft tissues in simple elongation," Am J Physiol 213(6), 1532-44 (1967).

[2] Fung, Y. C., [Biomechanics: Mechanical Properties of Living Tissues], New York: Springer, 568, (1993).

[3] Lee, S. H., Sifakis, E., and Terzopoulos, D., "Comprehensive biomechanical modeling and simulation of the upper body," ACM Siggraph (2010).

[4] Mayah, A. A., Moseley, J., Hunter, S., Velec, M., Chau, L., Breen, S., and Brock, K., "Biomechanically-based image registration of head and neck radiation treatment," Physics in Medicine and Biology 55, 6491-6500 (2010).

[5] Chadwick, J. E., Huamann, D. R., and Parent, R. E., "Layered constricution of deformable animated characters," Computer Graphics 23(3), 243-252 (1989).

[6] Kahler, K., Haber, J., Yamauchi, H., and Seidel, H., "Head shop: generating animated head models with anatomical structure," ACM Siggraph, 55-64 (2002).

[7] Sifakis, E., Neverov, I., and Fedkiw, R., "Automatic determination of facial muscle activations from sparse motion capture marker data," ACM Transactions on Graphics 24(3), 417-425 (2005).

[8] Lee, S. H. and Terzopoulos, D., "Heads Up! Biomechanical modeling and neuromuscular control of the neck," ACM Transactions on Graphics 25(3), 1188-1198 (2006).

[9] Zordan, V. B., Celly, B., Chiu, B., and Dilorenzo, P. C., "Breathe easy: model and control of simulated respiration for animation," ACM Siggraph, 29-37 (2004).

[10] Dilorenzo, P. C., Zordan, V. B., and Sanders, B. L., "Laughing out loud: Control for modeling anatomically inspired laughter using audio," ACM Transactions on Graphics 27(5), 125:1-8 (2008).

[11] Albrecht, I., Haber, J., and Seidel, H., "Construction and animation of anatomically based human hand models," ACM Siggraph, 98-109 (2003).

[12] Tsang, W., Singh, K., and Fiume, E., "Helping hand: An anatomically accurate inverse dynamics solution for unconstrained hand motion.," ACM Siggraph, 319-328 (2005).

[13] Van Nierop, O. A., Van Der Helm, A., Overbeeke, K. J., and Djajadiningrat, T. J., "A natural human hand model," The visual computer 24(1), 31-44 (2008).

[14] Sueda, S., Kauffman, A., and Pai, D. K., "Musculo tendon simulation of hand animation," ACM Transactions on Graphics 27(3), 831-838 (2008).

[15] Dong, F., Clapworthy, G., Krokos, M., and Yao, J., "An anatomy-based approach to human muscle modeling and deformation," IEEE Transactions on visualization and computer graphics 8(2), 154-170 (2002).

[16] Komura, T., Shinagawa, Y., and Kunii, T. L., "Creating and retargeting motion by the musculoskeletal human body model," The visual computer 16(5), 254-270 (2000).

[17] Santhanam, A., Physics in Medicine and Biology (2010).

[18] Santhanam, A., Willoughby, T., Kaya, I., Shah, A., Meeks, S. L., Rolland, J. P., and Kupelian, P., "A Display Framework for Visualizing Real-Time 3D Lung Tumor Radiotherapy," IEEE Journal of Display Technology, 473-482 (2008).

[19] Santhanam, A., Imielinska, C., Davenport, P., Kupelian, P., and Rolland, J., "Modeling and simulation of realtime 3D lung dynamics," IEEE Information Technology and Biomedicine 12(2), 257-270 (2006).

[20] Santhanam, A., Willoughby, T., Shah, A., Meeks, S. L., Rolland, J. P., and Kupelian, P., "Real-time Simulation of 4D lung tumor radiotherapy using a breathing model," Lecture Notes on Computer Science 11, 710-717 (2008).

[21] Picinbono, G., Delingette, H., and Ayache, N., "Non-linear anisotropic elasticity for real-time surgery simulation," Graphical Models 65, 305-321 (2003).

[22] Ogden, R. W., "Large Deformation Isotropic Elasticity: On the Correlation of Theory and Experiment for Compressible Rubberlike Solids," Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences 326(1567), 565-584 (1972).

[23] Korhonen, R. K. and Saarakkala, S., "Biomechanics and Modeling of Skeletal Soft Tissues," in [Theoretical Biomechanics], V. Klika, Editor, InTech (2011).

[24] Mooney, M., "A theory of large elastic deformation," Journal of Applied Physics 11(9), 582-592 (1940).

[25] Rivlin, R. S., "Large elastic deformations of isotropic materials. IV. Further developments of the general theory," Philisophical Transactions of the Royal Society of London. Series A: Mathematical and Physical Sciences 241(835), 379-397 (1948).

[26] Natali, A., Carniel, E., Pavan, P., Dario, P., and Izzo, I. "Hyperelastic models for the analysis of soft tissue mechanics: definition of constitutive parameters," in *Biomedical Robotics and Biomechatronics*, Pisa: IEEE (2006).

[27] Hibbit, Karlsson, and Sorensen, [ABAQUS theory manual], Pawtucket, RI, (1998).

[28] Suli, E. and Mayers, D., [An Introduction to Numerical Analysis], Cambridge, UK: Cambridge University Press, (2003).

[29] Neylon, J., Qi, X., Sheng, K., Staton, R., Pukala, J., Manon, R., Low, D. A., Kupelian, P., and Santhanam, A., "A GPU based high-resolution multilevel biomechanical head and neck model for validating deformable image registration," Med Phys 42(1), 232-43 (2015).

CHAPTER 8: Parameterized Image Similarity for Fast, Automated Clinical DIR Performance Assessment

A version of this chapter is currently being prepared for submission to Medical Physics

ABSTRACT

Purpose. Quantifying deformable image registration accuracy is a difficult task in a clinical setting due to poor image quality of the daily imaging modalities (CBCT, MVCT), and the lack of known ground truth deformations. The current gold standard is physical distance between manually placed landmarks, known as the target registration error. However, this process is time-consuming and subject to user-biases. Image similarity metrics (ISM) may provide an alternative way to represent the registration error, but need to be parameterized and translated to physical distance measures to enable a fast, quantitative comparison of registration performance.

Methods. Parameterization terms using ISMs were developed from two expectations of an accurate registration: (1) the warped data obtained from applying the deformation vector field (DVF) to the source data should closely match the target data, and (2) the similarity between the source and warped pair should match the similarity between the source and target pair.

A biomechanical model was used to create pairs of volumetric images with known ground-truth deformation vector fields. Numerous registrations were performed to systematically fill a four dimensional registration parameter space, in order to provide a full characterization of the relationship between TRE and the ISM terms. A cost function was then developed to test the relationship between the parameterized ISMs and the known registration error for sub-volumes enclosing critical radiotherapy structures in the head-and-neck region.

In order to translate the parameterized ISM terms into a physical error measure, the ground-truth data was also fed through a neural network, where a stochastic gradient descent algorithm was applied iteratively to optimize a non-linear model able to infer an estimate TRE.

Results. While the cost function showed that a relationship could be established between the image similarity metrics and target registration error, the function was not sophisticated enough to fully characterize the relationship. The trained neural network provided the necessary level of abstraction, and achieved 88% accuracy when trained on a systematic sampling of registration parameters for a single deformation, and over 95% accuracy when trained on a variety of different anatomies for a single patient. Additionally, correlations of 0.9 or better were achieved three of four contours investigated.

Conclusions. The formulation presented demonstrates the ability for fast, accurate quantification of registration performance. When sufficiently trained on annotated data, a neural network can learn to infer an expectation value of target registration error from parameterized image similarity metrics. Such networks have potential clinical impact in patient and site-specific optimization, and stream-lining clinical registration validation.

INTRODUCTION

Deformable image registration (DIR) has become an important tool in radiation therapy, allowing image-guided analyses of non-rigid anatomical variations [1, 2]. This has many clinical applications, including automatic contour propagation [3] and image-guided radiotherapy, and constitutes a major component of adaptive radiotherapy (ART) techniques [4]. Several recent studies have shown that ART can provide significant dosimetric benefits for inter-fraction anatomic variations, as well as reduced normal tissue toxicity, in the head-and-neck [5-8], as well as other cancer sites [9-12]. This is achieved by adapting the plan to a patient's daily anatomy, which may also allow a reduction in the error margins added around the clinical tumor volume (CTV) to construct the planning target volume (PTV) [13]. In order to adapt the plan, the delivered dose must be accumulated and mapped to the daily anatomy while the patient lies on the table in the treatment room, greatly shortening the time scales for registration and validation [14]. Clinical implementations of ART remain largely limited to off-line studies and require a significant amount of user intervention [5, 15]. The computational challenges and increased manpower requirements of ART has inhibited full on-line capabilities for daily monitoring of every patient[14]. For this methodology to be feasibly incorporated into the daily clinical workflow, speed and accuracy of DIR becomes paramount.

DIR Validation and Accuracy Quantification

The accuracy of DIR is critical to quantitatively track changes in patient anatomy, and the overall success of adaptive RT. Clinical DIR assessment has also been hampered by a lack of techniques to generate ground-truth deformations for evaluating and quantifying DIR performance. There has been much work in recent years assessing and comparing the accuracy of commercially available DIR algorithms [16-20]. However, these studies are performed offline and rely heavily on manually located landmarks to measure target registration error (TRE).

The availability of tools to quantitatively assess the accuracy of clinical registrations are lagging far behind the registration algorithms themselves [21].

Historically, registration accuracy has been measured by comparing the deformed image with the target image and measuring the difference between the estimated deformation and the true deformation, calculating the TRE. For clinical scenarios, however, the true deformation is unknown, so direct assessment is not possible without user intervention. Additionally, clinical registration accuracy is difficult to quantify due to poor image quality of the daily imaging modalities (CBCT, MVCT), and the lack of known ground truth deformations to validate the DIR algorithms. The current gold standard for obtaining the TRE requires comparison between manually placed corresponding landmarks on the source and target and calculating the difference between the user defined deformation and the deformation reported by the DIR [22]. However, placing landmarks is time intensive, subject to inter- and intra-observer variability, and suffers from small sample size [23, 24].

A fast automated methodology for assessing registration performance is necessary for implementation into the daily clinical workflow [14]. It has also been shown that registration performance can be improved by optimizing registration parameters on a per patient or per registration basis [25]. Furthermore, previous work has shown that registration performance is also site-specific [26, 27], demonstrating the need for narrow focus of DIR to the current clinical application. A fast automated methodology for assessing registration performance is necessary if patient or site specific registration optimization is ever going to be implemented into the daily clinical workflow.

Using Image Similarity Metrics to Assess DIR Performance

Image similarity metrics (ISMs) provide a fast method for assessing the correlation between two image sets and outputs a single value quantifying the similarity between intensity fields. However, the quantification has little meaning without a proper frame of reference [28]. Therefore, using image based metrics is currently qualitative, i.e. the range of values for each image based metric is not fixed. In 2012, Rohlfing presented an exhaustive study of the limitations of image similarity and tissue overlaps as accuracy measures for deformable image registration. He showed how direct application of these measures can be deceptive, reaffirming the lack of an automated method for quantifying DIR accuracy. He concluded that the gold standard remains manually placed landmarks, despite the inefficiency of the method and time requirement.

In this manuscript, we attempt to overcome these limitations by finding a mathematical relationship between the true landmark-based TRE and the ISM. The relationship was established by parameterizing the ISM and iterating over large correlated data sets. To contextualize the ISM, an initial cost function was proposed based on the expectation that for a good registration, the deformed image will behave similarly to the target image when processed in comparison to the source image. This should normalize the result relative to the initial similarity of the source and target, and enable comparisons between registrations on separate image sets.

Neural Network Approach Versus Manual Cost Function

While a cost function can provide good results by manually tweaking the parameterization variables to adjust the cost function response (CFR), this iterative method of fine-tuning can become just as time intensive as the process of manually placing landmarks. To further automate the process, a neural network was developed to find a non-linear system of equations able to predict the TRE from the similarity information.

Neural networks have gained significant traction in recent years for a wide variety of applications. Wu et al. have demonstrated the potential of a neural network based quality evaluator for rigid transformations during head-and-neck patient set up [29, 30].

The attraction lies in a neural network's ability to learn relationships from annotated data, without the necessity for user intervention to design specific features or identifiers. This is typically done using a form of stochastic gradient descent to modify weights and biases until the network output matches the expected results as closely as possible. The depth and scope of the network allows it to construct much more complex relationships. A trained network can then accurately infer a result from unlabeled input data.

Application of neural networking methods in the medical arena have been predominantly limited to the field of computer aided diagnosis [31-36]. The problem posed in this manuscript, however, does not require as many levels of abstraction compared to today's sophisticated image classification, object recognition, or segmentation networks. We hypothesize that the relationship between ISM and TRE can be modeled using a two-layer network with non-linear activation.

With a fully trained network, the framework should provide be able to provide a robust, quantified error expectation of DIR performance in near real-time. Such a fast and consistently reliable registration assessment has the potential to facilitate patient and site specific assessment and optimization.

MATERIALS and METHODS

Expectations of the ISM Response

The parameterization was developed around two expectations as the registration error approaches zero. The given image pairs are considered to be the source and target images. The warped dataset was created by applying the deformation vector field (DVF) obtained from the DIR algorithm to the source image. Similarity measures were calculated for three sets of images: source-target (I_{ST}), source-warp (I_{SW}), and target-warp (I_{TW}). The expectations can then be expressed as: (eq. 1) the similarity of the target and warped datasets should approach 1, and (eq. 2) the similarity between the source and warped datasets should approach the similarity between the source and target datasets.

$$Y = I_{TW} \tag{1}$$

$$X = 1 - |I_{ST} - I_{SW}| \tag{2}$$

For an ideal registration, X → 1, and Y → 1. The image similarity metric chosen for testing initial response was normalized mutual information (NMI), which uses the entropy of the individual images sets, H_A and H_B, and their combined entropy, H_{AB}; as shown in eq. 3 [37, 38].

$$NMI = \left(\frac{H_A + H_B}{H_{AB}}\right) - 1 \tag{3}$$

Generating Ground Truth Data

Simulated CTs with known DVFs from biomechanical modeling. In previous work, a framework was developed with the ability to instantiate an interactive biomechanical model from a patient CT [39]. These models were used to induce posture changes and physiological regression to simulate day-to-day changes in patient anatomy and create clinically realistic ground truth deformation vector fields for the purpose of clinical DIR validation. The framework outputs a simulated CT of the deformed anatomy and a fully volumetric DVF so the motion of each voxel was known.

Dense registration parameter space / TRE correspondence. An in-house multi-level, multi-resolution optical flow DIR algorithm was employed for these experiments [40]. The registration algorithm had four adjustable parameters: (1) the smoothing factor, (2) the number of resolution levels, (3) the number of iterations, and (4) the number warps. Registrations were performed for a systematic sampling of this four dimensional registration parameter space. Table 8.1 displays the sampling rate for each variable and the total number of registration performed. A total of 2400 registration were performed between the source-target dataset to create the dense parameter space data set. The induced deformation of the target image for this data set consisted of 15° rotations about each axis, and 25% regression in the primary tumor contour. This was a much larger change in anatomy than typically observed clinically, but was chosen to accentuate the differences For each registration, a deformed image volume was created from the DIR DVF, the ground-truth TRE (gt-TRE) calculated, and similarity analysis was run between the three sets of image pairs.

Table 8.1. Sampling frequency for each parameter of the 4D registration parameter space. A total of 2400 registrations were performed between a single source-target image set to create the dense parameter space data set.

Registration Parameter	Range	Instances
Warps	1:3	3
Levels	1:5	5
Iterations	50:500	8
Smoothing	10:1000	20
Total Registrations		2400

Additionally, annotated data was generated for a variety of anatomies by inducing 45 different postures with the biomechanical model, systematically rotating the head about the three primary axes. At each posture, 6 levels of regression were applied to the primary tumor target, creating a total of 270 target volumes with known deformations. Registrations were run between the source and each of these target volumes for 5 different smoothing parameters, and the gt-TRE was recorded by randomly selecting 100 landmarks within each structure of interest and comparing the DIR DVF with the known model DVF. Table 8.2 describes the composition of this multi-pose anatomy data set.

Table 8.2. Composition of annotated training data for systematic variations in head posture and tumor regression levels. Additionally, the smoothing parameter of the DIR algorithm was varied to create a total of 1350 registrations for the multi-pose anatomy data set.

	Levels of Regression	Postures			Registration Smoothing	Annotated Data Sets
Instances	6	45			5	1350
Range	0:30%	x-rotation -4:4	y-rotation -2:2	z-rotation -2:2	50:1000	

Sub-volume / site specific assessments. Analyzing the similarity of CT images of the head-and-neck at a full volumetric level can diminish the effectiveness due to the high percentage of empty space, and the lack of deformation in areas such as the brain and shoulders. Therefore, analysis was also performed on sub-volumes of the data. These sub-volumes were automatically generated with respect to the extents of the contoured structures of interest for radiotherapy purposes, including the right and left parotid glands, the spinal cord, and the tumor targets.

Establishing a Predictive Relationship Between ISMs and TRE

Manual cost function construction. Equation 4 shows the proposed cost function combining the similarity terms from equations (1) and (2), where *m* and *n* are variables to be optimized, and *f* is a weighting factor between 0 and 1. A systematic analysis was performed to determine the effect of the cost function variables (CFVs) (m,n,f) on the CFR. An inverse near-linear relationship was expected between the ISM values and the gt-TRE. By manipulating the CFVs, the response curve can be manipulated.

$$R = fX^m - (1-f)Y^n \tag{4}$$

Neural network construction. The proposed neural network was a fully connected two-layer network. As inputs, the values from equations 1 and 2 were calculated for the sub-volumes encompassing four critical structures in head-and-neck radiotherapy: the primary PTV, left parotid, right parotid, and cord. The output of the network was a vector of network predicted TRE (nn-TRE) values corresponding to the volumes encompassing each of these structures. The number of neurons in the hidden layer were optimized for the best result, ultimately settling at thirteen. A simple schematic of the network is shown in figure 8.1(a). Annotated data was split evenly between a training set and test set. As the training data is fed through the network, a series of weights and biases are adjusted to minimize a loss function. The accuracy of the network was continually monitored by inferring an output from the test data, and comparing to the ground truth expectations. Figure 8.1(b) illustrates the flow of data for the full network architecture. The eight input values are sent through the hidden and output layers, while the four known expectations are sent to the loss and accuracy functions. The result of the output layer is sent to the loss function, accuracy function, and training algorithm, which updates the weights and biases of the hidden and output layers and are used to calculate the network accuracy.

Neuron activation function. The neurons of the hidden layer take the data from each neuron of the input layer, apply a matrix multiplication with weighting factors, add a bias, and then apply a non-linear activation function. Without this activation function, the network would be comprised of linear function, shown in eq. 5, where the activation of a hidden neuron, a_j, is a linear function defined by the input, *z*, weight, *w*, and bias, *b*, summed over the input neurons, *i*. These weights and biases are the values adjusted during training and allows the network to learn.

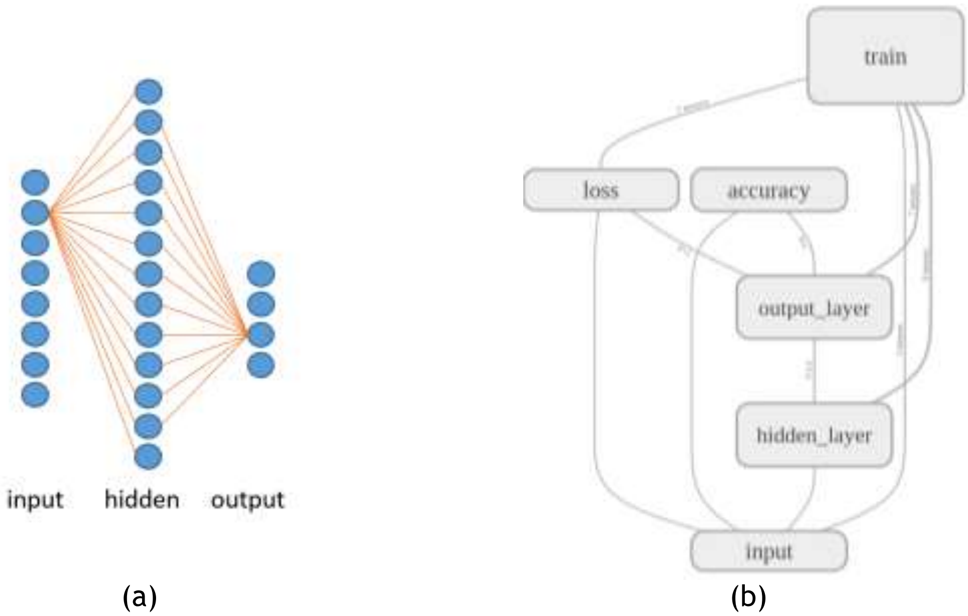


Figure 8.1. Neural network architecture. A conceptual representation of a three-layer fully connected network is shown in (a) with 8 input neurons, 13 hidden neurons, and 4 output neurons. A graph visualization of data flow for the network used in this manuscript is shown in (b), constructed using the TensorBoard graph visualization tool provided in the TensorFlow library.

$$a_j = \sum_i w_{j,i} z_i + b_j \tag{5}$$

Converting this to a non-linear response is important because a composition of linear functions remains a linear function, so the network abstraction is limited no matter its depth. The activation function chosen for this network was the sigmoid (eq. 6) [41].

$$\sigma(a) = \frac{1}{1+e^{-a}} \tag{6}$$

The sigmoid function is essentially a smoothed out step function. The sigmoid function was chosen because there was no loss of data for negative values, which is typical for other activation functions, such as hyperbolic tangent and rectified linear unit function. This was important because the network outputs a physical value. Sigmoid activation has the drawback of possibly saturating during training, but this was not much of a concern for the size of the network being employed in this manuscript.

Loss and accuracy measures. The loss function is applied during training to calculate the error between the output of the feed-forward network and the gt-TRE. Since the output of the network was intended to be a physical quantity, quadratic cost was implemented as the loss function [41, 42].

$$Loss = \frac{1}{N_s} \sum_s (\boldsymbol{y_s} - \boldsymbol{a_s})^2 \qquad (9)$$

Where N_s is the size of the training data set, s is the individual training data, and $\boldsymbol{y}$ was the tensor of true expected outcomes, and $\boldsymbol{a}$ was the tensor of network outputs.

The accuracy was calculated as an absolute percent error between the nn-TRE and the gt-TRE, with a target of 0.1 mm accuracy. Relative percent error had little physical meaning for instances where the gt-TRE approached zero. The denominator of 2 corresponds to an expected value of 2 mm for the gt-TRE, *y*. Therefore, 75% accuracy corresponded with a physical margin of 0.5 mm, and an accuracy of 0% corresponded with an error of 2 mm from the actual gt-TRE. Setting the expectation value at 2 mm matched the in-plane resolution of the CT data being registered, which had voxel dimensions of 1.953 mm by 1.953 mm with a slice thickness of 3 mm. Using this measure, anything less than 100% accuracy can be considered sub-voxel error.

$$Accuracy = \; 100 * \left(1 - \frac{y_s - a_s}{2}\right) \qquad (9)$$

The accuracy reported was also averaged over the number of samples in the test data during training, and the number of samples in the independent data when inferred results with the fully trained network.

Back-propagation. In order for the network to learn, the error from the loss function has to alter the network to better approximate the expected outcome. Backpropagation is a method of retracing the network from output to input, adjust weights, *w*, and biases, *b*, at each layer by applying their respective partial derivatives of the loss function [43]. This method became the prominent method for network training in 1986, when Rumelhart et al. showed the performance benefits utilizing gradient descent [44]. Currently, the most prominent method for neural network training is stochastic gradient descent [45]. Stochastic gradient descent (SGD) trains on smaller batches of training data, called epochs. Within an epoch, the training batch is iterated through several times, randomly choosing data points to estimate the gradient. An adaptive sub-gradient method, with dynamic learning rates was employed for network training [46].

Development Environment

The biomechanical model, registration algorithm, and image similarity analysis tools were developed on a Linux environment, using C/C++ and accelerated with NVIDIA's CUDA library to run on graphics processing units (GPUs). Neural network development was done in python, using the Google's open source library for machine intelligence, TensorFlow, accelerated for GPU with the CUDA deep neural network library, cuDNN.

RESULTS

Cost Function Response Versus Target Registration Error

Figure 8.2 shows the full 4D parameter space stretched over the x-axis, with plots of the gt-TRE and CFR in the primary and secondary y-axis, respectively, for the PTV1 volume. Within each level subdivision in the figure, there are 8 peaks corresponding to the varying number of iterations. The smooth curves between peaks correspond to the sampling of the smoothing parameter. As expected, there was an observable inverse correlation. While it appears that the moving average of the CFR increased in concordance with the moving average of the gt-TRE

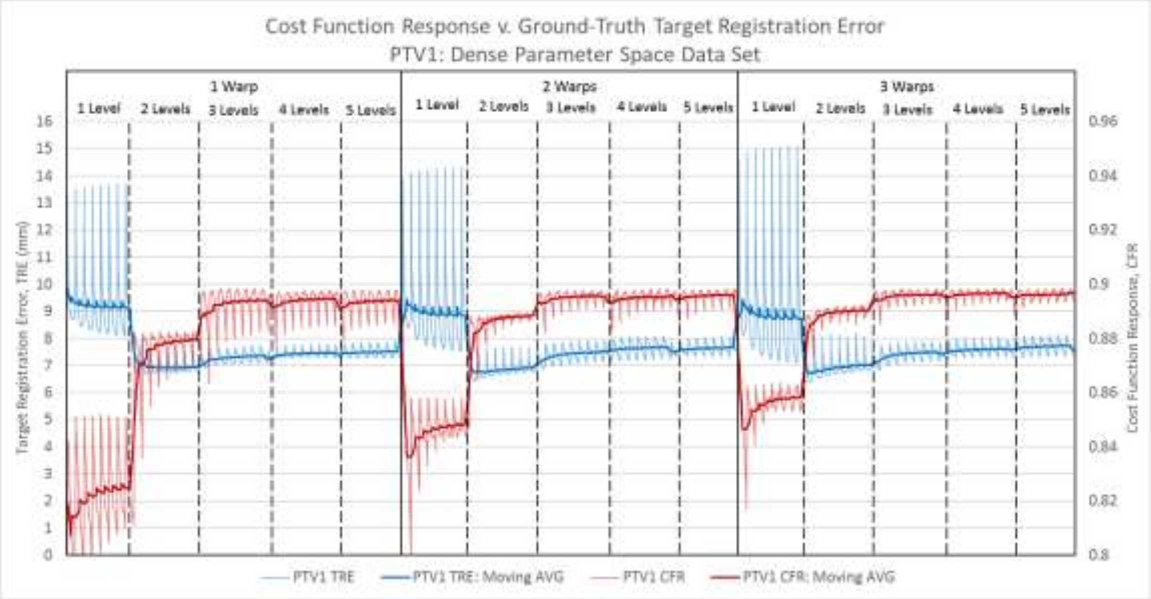


Figure 8.2. Comparison of TRE and CFR over the entire dense parameter space data set for the PTV1 contour, with moving averages for a window size of 20 samples. Four registration parameters were systematically sampled. Number of warps comprised the outer most loop, followed by number of levels, iterations, and smoothing value. The plot shows delineations of how the 4D parameter space was plotted in 1D along the x-axis for the warps and levels. An inverse correlation can be observed between TRE and CFR throughout the entire registration parameter space.

decreasing, the actual correlation between the data was just -0.4675 for the cost function variables used to generate the data in figure 8.2. When plotted against each other, shown in figure 8.3, the short-comings of the cost function become obvious.

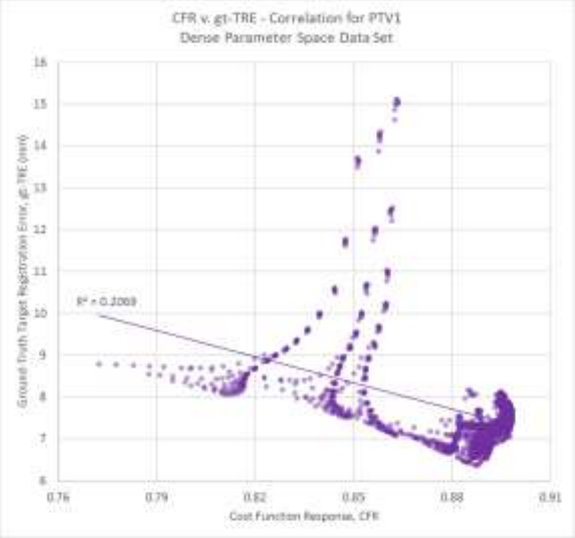


Figure 8.3. Plotting the cost function response against the target registration error for the PTV1 contour. The data plotted here corresponds to the data plotted in figure 8.1.

The cost function was able to establish a general trend of decreasing response for larger TRE values, but the lack of a one-to-one correspondence suggest the cost function was not sophisticated enough to capture the full relationship between TRE and image similarity metrics. The CFR also showed strong dependence on the CFV chosen.

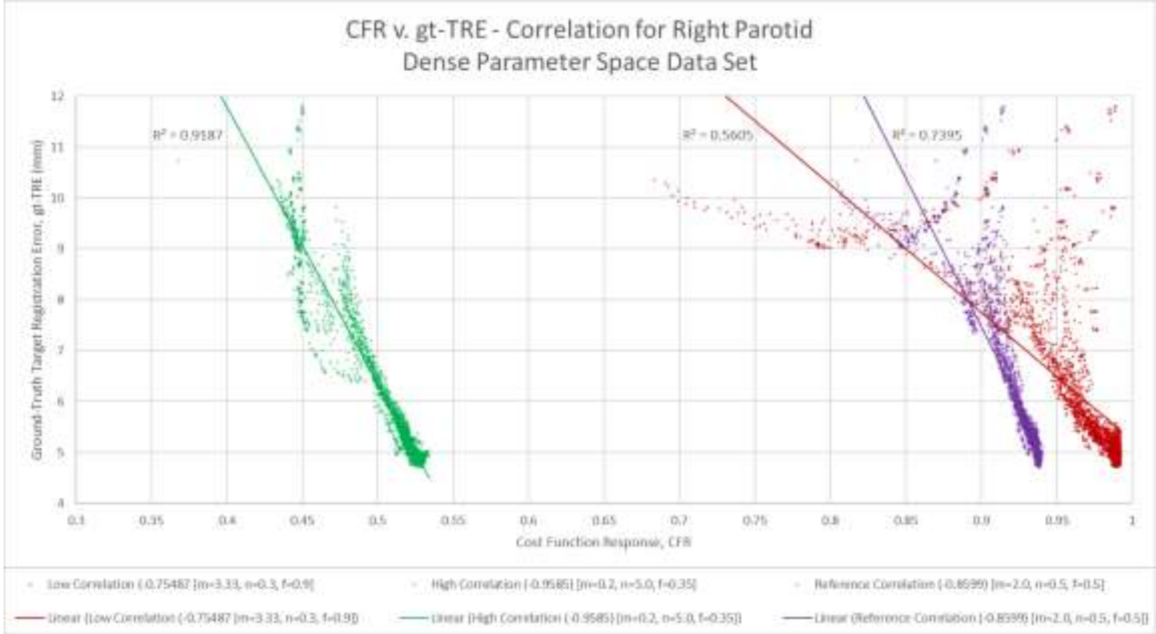


Figure 8.4. Correlation between target registration error (TRE) and cost function response (CFR) for three sets of cost function variables (CFV) using the full registration parameter space data, illustrating the high variability of response observed by adjusting the CFV.

For the data in figure 8.2 and 8.3, the weighting factor, *f*, was set to 0.5, and the exponents, *m* and *n*, were set as 2 and 0.5, respectively. This CFV set will be referred to as the reference CFV from this point on. Figure 8.4 shows how the shape of the CFR response can vary with the CFV. The data presented comes from the sub-volume surrounding the right parotid gland. For the right parotid, a high correlation was found (-0.9187), by adjusting the CFV. The best result CFV was also observed to be site-specific, and expected to be deformation-specific, and patient-specific as well. This increases the complexity of task-specific registration optimization. In the next section, the results for the neural network are presented, where manually adjusted variables are eliminated and replaced with a framework for learning from annotated training data.

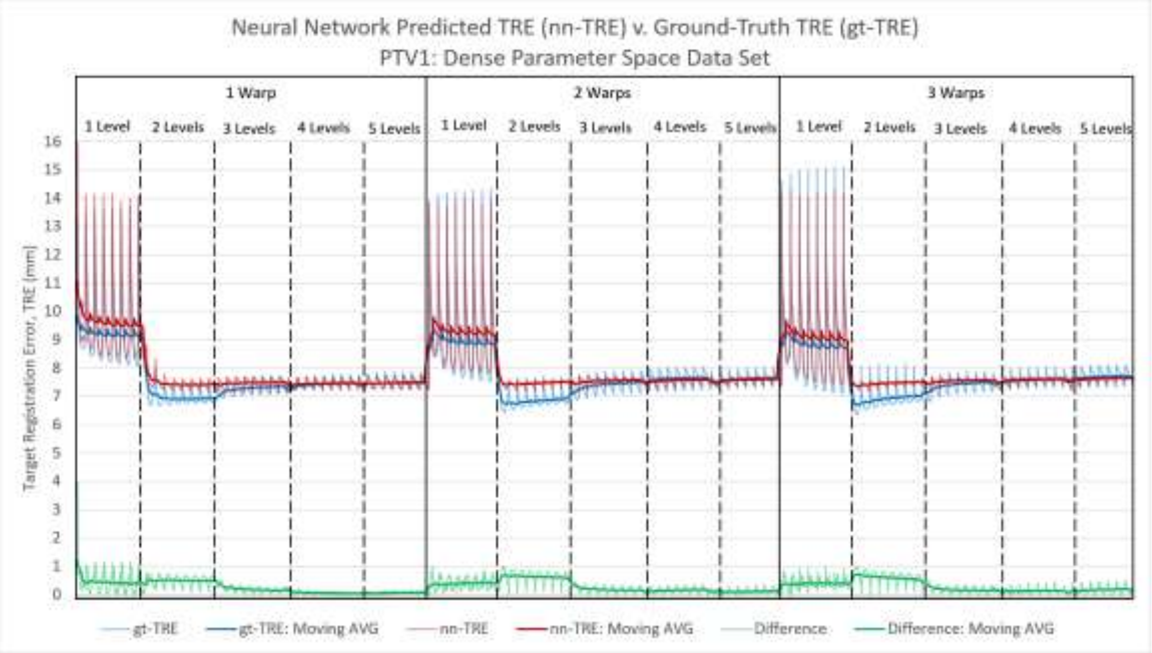


Figure 8.5. Comparison of gt-TRE and nn-TRE over the entire dense parameter space data set for the PTV1 contour, with moving averages for a window size of 20 samples. The neural network was able to predict the TRE to within 1 mm after being trained on 25% of the annotated ground truth data. The difference in mm is also plotted with its moving average.

Neural Network Results

Training on the dense parameter space data set. The results of the cost function indicated that a substantive relationship did exist between the proposed ISM expectation terms, *X* and *Y*, as described by equations 1 and 2, but that a more complex parameterization was necessary to fully characterize that relationship. A neural network was developed which took as inputs the calculated similarity metrics, *X* and *Y*, and would infer a predicted target registration error (nn-TRE) as an output. From the 2400 samples in the Dense Parameters Space data set, 25% or 600 were chosen randomly as the training data. The network was trained in batches of 75 samples over 1000 epochs, such that every 8 epochs, all 600 training data had been iterated through. After training, the entire data set was fed through the network. The network reached 88% accuracy on the test data, which consisted of 75% of the dense parameter space data set. The results are shown in figure 8.5, along with the difference in millimeters between the predicted TRE and the true TRE.

Table 8.3. Correlation with ground-truth TRE for cost function response before and after optimizing the cost function variables, and for the neural network predicted TRE, trained on 25% of the Dense Parameter Space data set.

	CFR v. gt-TRE		nn-TRE v. gt-TRE
	Reference CFV	Best CFV	
PTV1	-0.467	-0.649	0.950
Left Parotid	-0.921	-0.952	0.988
Right Parotid	-0.860	-0.958	0.952
Cord	0.138	-0.109	0.753

The network was able to predict the TRE to within 1 mm for the entire registration parameter space, excluding the purposely poor registrations performed using only 1 warp, 1 level, and a less than 100 iterations. The mean accuracy for the PTV1 was just under 86%, corresponding to a mean discrepancy less than 0.3 mm. The correlation between the nn-TRE and gt-TRE matched or exceeded the best performance from manual optimization of the cost function for each of the four contours examined. These results are shown in table 8.3.

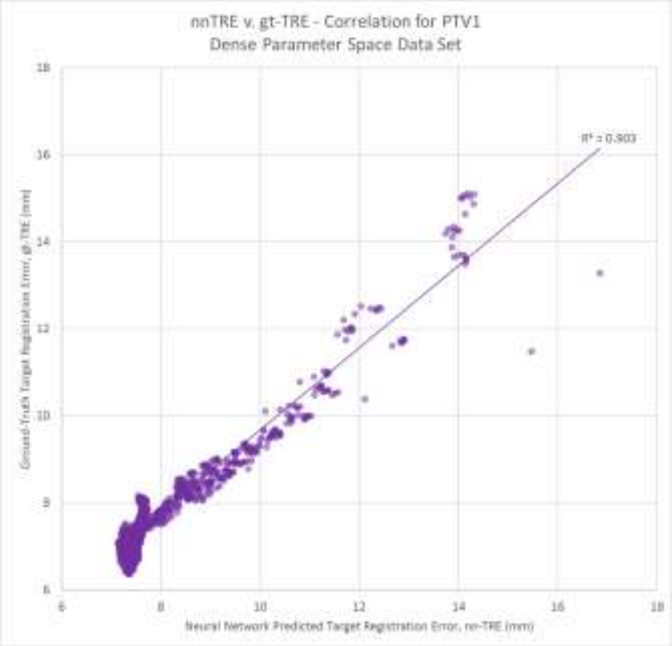


Figure 8.6. Plotting the nn-TRE against gt-TRE of the PTV1 contour for the dense parameter space data set.

Figure 8.6 plots the nn-TRE with respect to the gt-TRE, in comparison to figure 8.3, showing much better correspondence with a tight grouping along the linear trend-line. Results show the neural network can accurately predict the TRE for a large range of registration parameter combinations and the resultant range in registration quality. This shows potential for automated registration optimization, with the level of specificity (patient, site) determined by the annotated training data.

Training on the multi-pose anatomy data set. The first experiment tested whether the neural network could be trained to identify the best set of registration parameters for a single deformation. The second experiment was developed to test whether the neural network could be trained to predict the registration error for a variety of different anatomies that could be seen from day-to-day in the clinic. The multi-pose anatomy data set consisted of 45 different postures, and applied 6 levels of tumor regression at each posture. In addition to that, registrations were performed for 5 different smoothing values ranging from 50 to 1000. The smoothing variable dictates the scope of local continuity for the deformation vector field. The other registration parameters were set to constant values that are reasonable for clinical registrations. Therefore, the multi-pose anatomy data set consisted of clinically realistic day-to-day anatomies, with relatively small deformations, where DIR performance is expected to be good. For each pose, only subtle differences were expected between registrations based on the different smoothing parameters.

The architecture of the neural network remained the same for both experiments. The network was re-trained on 25% of the multi-pose anatomy data set, and achieved over 95% accuracy on the test data. The results for the PTV1 contour are shown in figure 8.7. It is apparent from the figure that the registrations as a whole were much better than the registrations in the dense parameter space data set, with the moving average of the gt-TRE ranging from 0.6 and 1.3 mm. However, there was still a large amount of variation between registrations, and the neural network was able to accurately reproduce the high frequency fluctuations with an average error of less than 0.1 mm. The only instance of significant deviation between the nn-TRE and the gt-TRE was for the case of no deformation seen in the middle of the 0% regression block. The

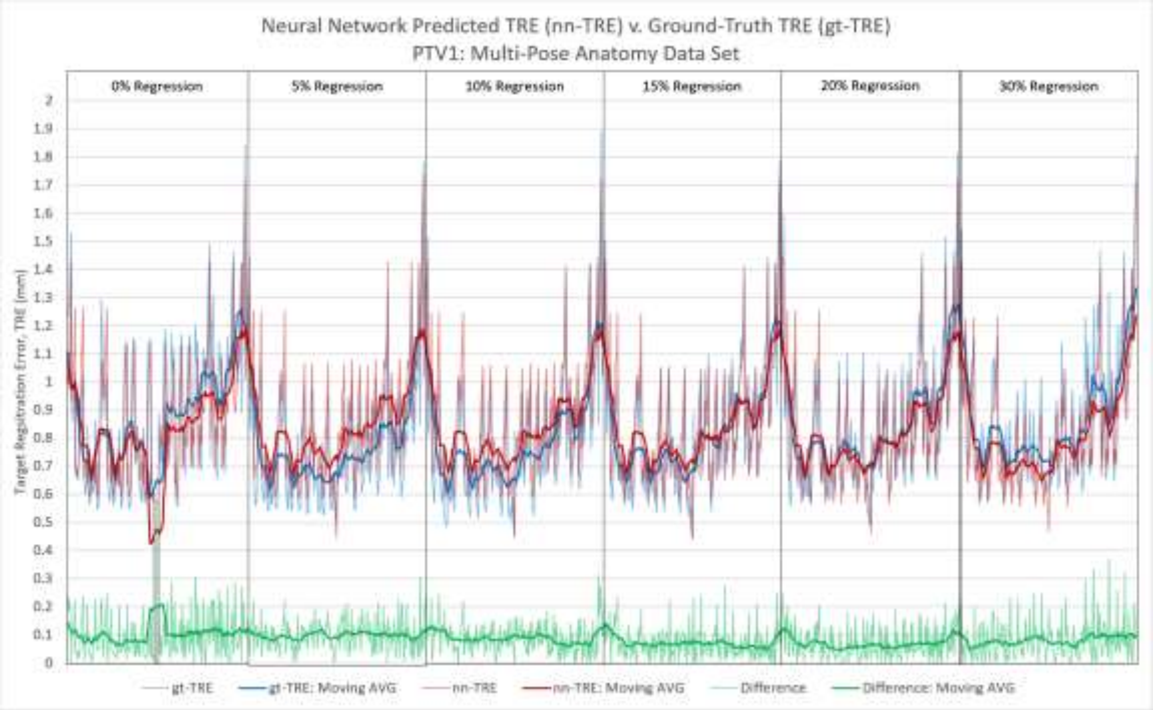


Figure 8.7. Comparison of gt-TRE and nn-TRE over the entire multi-pose anatomy data set for the PTV1 contour, with moving averages for a window size of 20 samples. The neural network maintained an average error of approximately 0.05 mm after being trained on 25% of the annotated ground truth data. The difference in mm is also plotted with its moving average. The segmentation of regression level are shown on the figure. Within each regression segment are 45 different postures, each of which was registered 5 times with different smoothing parameters.

correlation between the nn-TRE and gt-TRE for the multi-pose anatomy data set was 0.889 for the PTV1, 0.95 for the right parotid, and 0.945 for the left parotid.

This result supports the hypothesis that the neural network can reliably infer the TRE from only image similarity information for patient-specific scenarios, when properly trained using annotated data. The considerations for size and scope of the training data, as well as the potential avenues for incorporating a neural network are discussed below.

DISCUSSION

Dependence on Annotated Training Data

While neural networks appear to have great potential for fast, quantitative estimates of DIR performance, they are completely dependent on the accuracy and reliability of their annotated training data. Tables 8.4 and 8.5 report the accuracy results for the nn-TRE for different compositions and distributions of the training data for the dense parameter space data set and the multi-pose anatomy data set, respectively. Results for the dense parameter space data set showed better results when the training data was randomly sampled from the entire data set, as opposed to training on the entirety of a sequential third of the data. When training on the randomly sampled data, little difference was seen when the amount of data allocated for training was increased from 25% to 50% or 75%. This result held true for the multi-pose anatomy data set as well.

Table 8.4. Accuracy of the neural network predicted TRE on the dense parameter space data set for different compositions of the annotated training data

Training Data Composition	25% Random		50% Random		75% Random		First Third		Last Third	
	%	mm	%	mm	%	mm	%	mm	%	mm
PTV1	85.67	0.29	87.77	0.24	88.35	0.23	78.45	0.43	84.78	0.30
Left Parotid	81.90	0.36	81.01	0.38	79.89	0.40	78.92	0.42	79.90	0.40
Right Parotid	91.61	0.17	92.24	0.16	91.54	0.17	89.01	0.22	89.80	0.20
Cord	92.84	0.14	92.77	0.14	92.57	0.15	92.23	0.15	92.41	0.15
Overall	88.01	0.24	88.45	0.23	88.09	0.24	84.67	0.31	86.72	0.27

The experiments in this manuscript were limited to two separate-but-related, patient-specific tasks. Still the network trained on the dense parameter space data set was useless for predictions on the multi-pose anatomy data set, and vice versa. It would be feasible to generate a suite of postures like the multi-pose anatomy data set, for each patient in the clinic. Generating the suite of different postures, and running the registrations to produce the annotated TRE data took approximately one full day using a GPU-based biomechanical model and a fast optical flow deformable image registration algorithm. Network training over 1000 epochs took only a couple minutes. Once trained, the network can infer an estimated TRE almost instantaneously, needing only to wait for the image similarity to be calculated.

Table 8.5. Accuracy of the neural network inferred TRE on the multi-pose anatomy data set for different compositions of the annotated training data

Training Data Composition	25% Random		50% Random		75% Random	
	%	mm	%	mm	%	mm
PTV1	95.66	0.09	95.72	0.09	95.71	0.09
Left Parotid	92.31	0.15	91.08	0.18	91.20	0.18
Right Parotid	93.27	0.13	93.06	0.14	93.11	0.14
Cord	99.78	0.00	99.74	0.01	99.70	0.01
Overall	95.26	0.09	94.90	0.10	94.93	0.10

Further investigation is needed to determine how broad of a scope the network can have while maintaining accuracy. We focused on analyzing four critical structures for head-and-neck radiotherapy simultaneously. Any increase in scope would bring an accompanying requirement for more training data and additional network complexity in the form of more hidden layers or more neurons per layer. Similarly, training individual networks for each structure may improve network results and decrease the amount of training data required. Determining where and how to apply such networks should be an intense area of research, as their applications are wide-ranging and largely unexplored. The inhibiting factor will most likely remain the time and effort required to compile annotated training data, which further highlights how a fast, versatile, and accurate biomechanical model can be an invaluable resource.

Improving Registration Performance

This manuscript focused on quantifying the expected error in the deformable image registration, but a similar methodology could be applied for registration optimization. This again delves into the pre-determination of how and when the neural network should be employed. It may be possible to train a network to choose the best combination of registration parameters based on image similarity analysis of the source and target images. Once registered, a second network would give a quantified confidence of the registration performance. Alternatively, the network predicted TRE could be incorporated into a feedback loop with the registration algorithm for task or site-specific optimization.

Contour specific results were calculated in this manuscript by analyzing the sub-volumes encompassing them. This provided more information than a single measure of similarity between the entire 3D data volumes, and limited focus to the areas of greatest interest. Future work will investigate the calculation of a volumetric image similarity using a moving window throughout the entire 3D image set, similar to the application of a convolution filter. Greater weight can still be given to contours of interest, while delivering more detailed information. Coupled with a neural network, this could produce a volumetric measure of the DIR confidence, shown as a heat map to identify problem areas or to adapt the registration to the clinical task. Additionally, this could lead to adaptive registration parameters such as heterogeneous smoothing values.

CONCLUSION

The work in this manuscript delivers proof-of-concept that a neural network can infer an estimated TRE from image similarity information when properly trained with annotated data. For reasonable registrations applied to a variety of possible daily deformations, the network achieved greater than 95% accuracy when comparing its inferred TRE to ground-truth TRE. Once trained, the network requires only image similarity information in order to provide a confidence measure of the registration in real-time. Neural networks have the potential to be developed into a fully automated methodology for quantifying registration performance.

REFERENCES

[1] Sotiras, A., Davatzikos, C., and Paragios, N., "Deformable medical image registration: A survey," IEEE Trans Med Imaging 32(7), 1153-90 (2013).

[2] Crum, W. R., Hartkens, T., and Hill, D. L., "Non-rigid image registration: Theory and practice," Br J Radiol 77 Spec No 2, S140-53 (2004).

[3] Hardcastle, N., van Elmpt, W., De Ruysscher, D., Bzdusek, K., and Tome, W. A., "Accuracy of deformable image registration for contour propagation in adaptive lung radiotherapy," Radiation Oncology 8, 243 (2013).

[4] Yan, D., Vicini, F., Wong, J., and Martinez, A., "Adaptive radiation therapy," Physics in Medicine and Biology 42(1), 123-32 (1997).

[5] Capelle, L., Mackenzie, M., Field, C., Parliament, M., Ghosh, S., and Scrimger, R., "Adaptive radiotherapy using helical tomotherapy for head-and-neck cancer in definitive and postoperative settings: Initial results," Clinical Oncology 24(3), 208-215 (2012).

[6] Foroudi, F., Wong, J., Kron, T., Rolfo, A., Haworth, A., Roxby, P., Thomas, J., Herschtal, A., Pham, D., Williams, S., Tai, K. H., and Duchesne, G., "Online adaptive radiotherapy for muscle-invasive bladder cancer: Results of a pilot study," International Journal of Radiation Oncology Biology Physics 81(3), 765-771 (2011).

[7] Zeidan, O. and Huddleston, A. J., "A comparison of soft-tissue implanted markers and bony anatomy alignments for image-guided treatments of head and neck cancers.," International Journal of Radiation Oncology Biology Physics (2009).

[8] Zeidan, O. and Langen, K. M., "Evaluation of image-guidance protocols in the treatment of head and neck cancers," International Journal of Radiation Oncology Biology Physics 67(3), 670-677 (2007).

[9] Lindegaard, J., Fokdal, L., Nielsen, S., Juul-Christensen, J., and Tanderup, K., "Mri-guided adaptive radiotherapy in locally advanced cervical cancer from a nordic perspective," Acta Oncologica 52(7), 1510-1519 (2013).

[10] Nijkamp, J., Marijnen, C., Herk, M. V., Triest, B. V., and Sonke, J., "Adaptive radiotherapy for long course neo-adjuvant treatment of rectal cancer," Radiotherapy and Oncology 103(3), 353-359 (2012).

[11] Schwartz, D., Garden, A., Thomas, J., Chen, Y., Zhang, Y., Lewin, J., Chambers, M., and Dong, L., "Adaptive radiotherapy for head-and-neck cancer: Initial clinical outcomes from a prospective trial," International Journal of Radiation Oncology Biology Physics 83(3), 986-993 (2012).

[12] Tuomikoski, L., Collan, J., Keyrilainen, J., Visapaa, H., Saarilahti, K., and Tenhunen, M., "Adaptive radiotherapy in muscle invasive urinary bladder cancer - an effective method to reduce the irradiated bowel volume," Radiotherapy and Oncology 99(1), 61-66 (2011).

[13] Qi, X. S., Neylon, J., Can, S., Staton, R., Pukala, J., Kupelian, P., and Santhanam, A., "Feasibility of margin reduction for level ii and iii planning target volume in head-and-neck image-guided radiotherapy - dosimetric assessment via a deformable image registration framework," Current Cancer Therapy Reviews 10(4), 323-333 (2014).

[14] Xing, L., Siebers, J., and Keall, P., "Computational challenges for image-guided radiation therapy: Framework and current research," Seminars in Radiation Oncology 17(4), 245-257 (2007).

[15] Veiga, C., McClelland, J., Moinuddin, S., Lourenco, A., Ricketts, K., Annkah, J., Modat, M., Ourselin, S., D'Souza, D., and Royle, G., "Toward adaptive radiotherapy for head and neck patients: Feasibility study on using ct-to-cbct deformable registration for "dose of the day" calculations," Medical Physics 41(3), 031703 (2014).

[16] Brock, K. K. and Deformable Registration Accuracy, C., "Results of a multi-institution deformable registration accuracy study (midras)," International Journal of Radiation Oncology Biology Physics 76(2), 583-96 (2010).

[17] Fabri, D., Zambrano, V., Bhatia, A., Furtado, H., Bergmann, H., Stock, M., Bloch, C., Lutgendorf-Caucig, C., Pawiro, S., Georg, D., Birkfellner, W., and Figl, M., "A quantitative comparison of the performance of three deformable registration algorithms in radiotherapy," Z Med Phys 23(4), 279-90 (2013).

[18] Hoffmann, C., Krause, S., Stoiber, E. M., Mohr, A., Rieken, S., Schramm, O., Debus, J., Sterzing, F., Bendl, R., and Giske, K., "Accuracy quantification of a deformable image registration tool applied in a clinical setting," J Appl Clin Med Phys 15(1), 4564 (2014).

[19] Mencarelli, A., van Kranen, S. R., Hamming-Vrieze, O., van Beek, S., Nico Rasch, C. R., van Herk, M., and Sonke, J. J., "Deformable image registration for adaptive radiation therapy of head and neck cancer: Accuracy and precision in the presence of tumor changes," International Journal of Radiation Oncology Biology Physics 90(3), 680-7 (2014).

[20] Varadhan, R., Karangelis, G., Krishnan, K., and Hui, S., "A framework for deformable image registration validation in radiotherapy clinical applications," J Appl Clin Med Phys 14(1), 4066 (2013).

[21] van Rijssel, M. J., Dahele, M., Verbakel, W. F., and Rosario, T. S., "A critical approach to the clinical use of deformable image registration software. In response to meijneke et al," Radiotherapy and Oncology 112(3), 447-8 (2014).

[22] Fitzpatrick, J. M. and West, J. B., "The distribution of target registration error in rigid-body point-based registration," IEEE Trans Med Imaging 20(9), 917-27 (2001).

[23] Woods, R. P., Grafton, S. T., Holmes, C. J., Cherry, S. R., and Mazziotta, J. C., "Automated image registration: I. General methods and intrasubject, intramodality validation," J Comput Assist Tomogr 22(1), 139-52 (1998).

[24] Woods, R. P., Grafton, S. T., Watson, J. D., Sicotte, N. L., and Mazziotta, J. C., "Automated image registration: Ii. Intersubject validation of linear and nonlinear models," J Comput Assist Tomogr 22(1), 153-65 (1998).

[25] Stanley, N., Glide-Hurst, C., Kim, J., Adams, J., Li, S., Wen, N., Chetty, I. J., and Zhong, H., "Using patient-specific phantoms to evaluate deformable image registration algorithms for adaptive radiation therapy," J Appl Clin Med Phys 14(6), 4363 (2013).

[26] Kirby, N., Chuang, C., Ueda, U., and Pouliot, J., "The need for application-based adaptation of deformable image registration," Medical Physics 40(1), 011702 (2013).

[27] Nie, K., Chuang, C., Kirby, N., Braunstein, S., and Pouliot, J., "Site-specific deformable imaging registration algorithm selection using patient-based simulated deformations," Medical Physics 40(4), 041911 (2013).

[28] Crum, W. R., Hill, D. L., and Hawkes, D. J., "Information theoretic similarity measures in non-rigid registration," Inf Process Med Imaging 18, 378-87 (2003).

[29] Wu, J. and Murphy, M. J., "A neural network based 3d/3d image registration quality evaluator for the head-and-neck patient setup in the absence of a ground truth," Medical Physics 37(11), 5756-64 (2010).

[30] Wu, J., Su, Z., and Li, Z., "A neural network-based 2d/3d image registration quality evaluator for pediatric patient setup in external beam radiotherapy," J Appl Clin Med Phys 17(1), 5235 (2016).

[31] Fonseca, P., Mendoza, J., Wainer, J., Ferrer, J., Pinto, J., Guerrero, J., and Castaneda, B., "Automatic breast density classification using a convolutional neural network

architecture search procedure," Medical Imaging, Computer Aided Diagnosis, SPIE 9414 (2015).

[32] Cruz-Roa, A., Basavanhally, A., Gonzalez, F., Gilmore, H., Feldman, H., Ganesan, S., Shih, N., Tomaszewski, J., and Madabhushi, A., "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks," Medical Imaging, Digital Pathology, SPIE 9041 (2014).

[33] Wang, H., Cruz-Roa, A., Basavanhally, A., Gilmore, H., Shih, N., Feldman, M., Tomaszewski, J., Gonzalez, F., and Madabhushi, A., "Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features," Journal of Medical Imaging 1(3) (2014).

[34] Bar, Y., Diamant, I., Wolf, L., and Greenspan, H., "Deep learning with non-medical training used for chest pathology identification," Medical Imaging, Computer Aided Diagnosis, SPIE 9414 (2015).

[35] Roth, H., Lu, L., Seff, A., Cherry, K., Hoffman, J., Wang, S., Liu, J., Turkbey, E., and Summers, R., "A new 2.5d representation for lymph node detection using random sets of deep convolutional neural network observations," Lecture Notes in Computer Science, MICCAI 8673, 520-527 (2014).

[36] Roth, H., Yao, J., Lu, L., Stieger, J., Burns, J., and Summers, R., "Detection of sclerotic spine metastases via random aggregation of deep convolutional neural network classifications," arXiv (2014).

[37] Roche, A., Malandain, G., Pennec, X., and Avache, N., "The correlation ratio as a new similarity measure for multimodal image registration," Lecture Notes on Computer Science 1496(MICCAI'98), 1115-1124 (1998).

[38] Wachowiak, M., Smolikova, R., and Peters, T., "Multiresolution biomedical image registration using generalized information measures," Lecture Notes on Computer Science 2879(MICCAI'03), 846-853 (2003).

[39] Neylon, J., Qi, X., Sheng, K., Staton, R., Pukala, J., Manon, R., Low, D. A., Kupelian, P., and Santhanam, A., "A gpu based high-resolution multilevel biomechanical head and neck model for validating deformable image registration," Medical Physics 42(1), 232-43 (2015).

[40] Min, Y., Neylon, J., Shah, A., Meeks, S., Lee, P., Kupelian, P., and Santhanam, A., "4d-ct lung registration using anatomy-based multi-level multi-resolution optical flow analysis and thin-plate splines," International Journal of Computer Assisted Radiology and Surgery 9(5), 875-889 (2014).

[41] Neilsen, M. *Neural networks and deep learning*. 2015.

[42] Kline, D. and Berardi, V., "Revisiting squared-error and cross-entropy functions for training neural network classifiers," Neural Computing & Applications 14(4), 310-318 (2005).

[43] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P., "Gradient-based learning applied to document recognition," Proceedings of the IEEE (1998).

[44] Rumelhart, D., Hinton, G., and Williams, R., "Learning internal representations by error propagation," Parallel distributing processing: Exploration in the microstructure of cognition, 318-362 (1986).

[45] Bottou, L., "Large-scale machine learning with stochastic gradient descent," Proceedings of COMPSTAT, 177-186 (2010).

[46] Zeiler, M. D., "Adadelta: An adaptive learning rate method," arXiv:1212.5701v1 (2012).

CHAPTER 9 – Conclusion of the Dissertation

Summary of Work

The projects discussed in this dissertation present solutions for the major inhibiting factors of time and reliability for several components of on-line ART, including dose calculation, dose accumulation, deformable image registration validation, verification, and error quantification. Algorithms were developed to automate several time-intensive tasks in the adaptive therapy workflow, and were accelerated to near real-time performance through parallelization and optimization for GPU architecture. These works ultimately provide tools for the radiotherapy community to move from conventional radiotherapy towards on-line adaptive therapy.

Chapter 2 describes the development of a highly parallelized, non-voxel-based (NVB) dose convolution algorithm developed and optimized for the unique memory hierarchy of GPU architecture. The NVB convolution method yielded total performance improvement factor of >4,000 when compared to a voxel-based ground truth CPU benchmark, and a factor of 20 compared with a voxel-based GPU dose convolution method. Accuracy compared to a CPU-computed ground truth dose distribution maintained greater than 99% of voxels passing the gamma test with 1% and 1 mm criteria.

In chapter 3, the NVB dose algorithm was redesigned to scale across multiple GPUs and an arbitrary number of machines in a multi-GPU cloud-based server (MGCS) framework. Despite the additional overhead of instantiating a distributed solution and transferring memory between server nodes, the MGCS implementation was able to achieve 2x performance for a simple, square field calculation when compared to a local, single-GPU implementation. This provides even greater computing power for maintaining sub-second calculations speeds for even the most complicated treatment plans and plan optimization methodologies.

In chapter 4, a GPU-accelerated DIR framework was developed, which utilized a fast optical flow DIR algorithm, to track changes in anatomy by registering the planning CT with the daily imaging and compare the planned dose distribution with the accumulated delivered dose. The framework incorporated several GPU-accelerated tools for contour-specific reporting, including Jacobian and Gamma analysis, and DVH generation. Using this framework, a retrospective study using weekly diagnostic CT scans showed dramatic volume changes and dosimetric deviations, including a minimum dose to the target up to 15% lower than the plan, and mean dose to the parotid that were significantly higher.

In chapter 5, a retrospective study using the DIR and dose accumulation framework showed that error margins could be reduced, improving normal tissue sparing while maintaining adequate tumor coverage. Reductions over 7% were seen in the maximum dose to the spinal cord, and over 18% in the mean dose to the parotids, while maintaining acceptable target coverage. These retrospective studies illustrate the capabilities of the dose accumulation framework, but also underline the potential benefit of incorporating daily online ART into the standard clinical workflow. However, the framework was dependent on the accuracy of its DIR algorithm, so a methodology was developed for validation on clinically realistic deformations.

This methodology was detailed in chapter 6, where a framework was developed to generate high-resolution, patient-specific, biomechanical models. The model performed at interactive speeds (>30 fps) on a single GPU for a model with approximately 1.5 million masses and over 25 million springs. The deformation of the HN anatomy by the model agreed with the clinically observed deformations with an average correlation coefficient of 0.956. The biomechanical model simulated a spectrum of posture and physiological changes, creating 270 unique anatomies for each of 10 clinical patient CT scans and outputting ground truth deformation vector fields that could be used to quantify DIR performance in a fully automated process. The amount of data production and analysis would have been insurmountable without GPU acceleration.

In chapter 7, a more sophisticated, hyper-elastic material model was incorporated into the biomechanical modelling framework to better simulate soft tissue deformations. The hyper-elasticity and fast re-meshing algorithm provided stability and accuracy to simulate larger deformations, and expand the applicability of the model beyond the head-and-neck site.

While the model provided an automated methodology for initial validation of DIR algorithms, quantifying the accuracy of clinical registrations remained a time-consuming, manual task in the on-line ART workflow. Chapter 8 presented the development of a neural network, able to infer the target registration error for sub-volumes around critical radiotherapy structures using parameterized image similarity metrics. The neural network achieved sub-millimeter accuracy compared to ground-truth model-generated deformations for two separate datasets examining a variety of registration parameters and clinically realistic deformations.

All these tools heavily utilize GPUs to accelerate computations to near real-time performance, leaving one final obstacle on the road to clinical implementation: integration. In the next section, future directions for each of the projects are discussed, including how they may be integrated together.

Future Directions

Figure 9.1 revisits the potential workflow for on-line ART of the introduction, where each of the projects discussed in this dissertation have been ported to distributed MGCS implementations. I envision the development of a web-based front-end user interface for interaction and visualization, with the full suite of GPU-accelerated tools running back-end

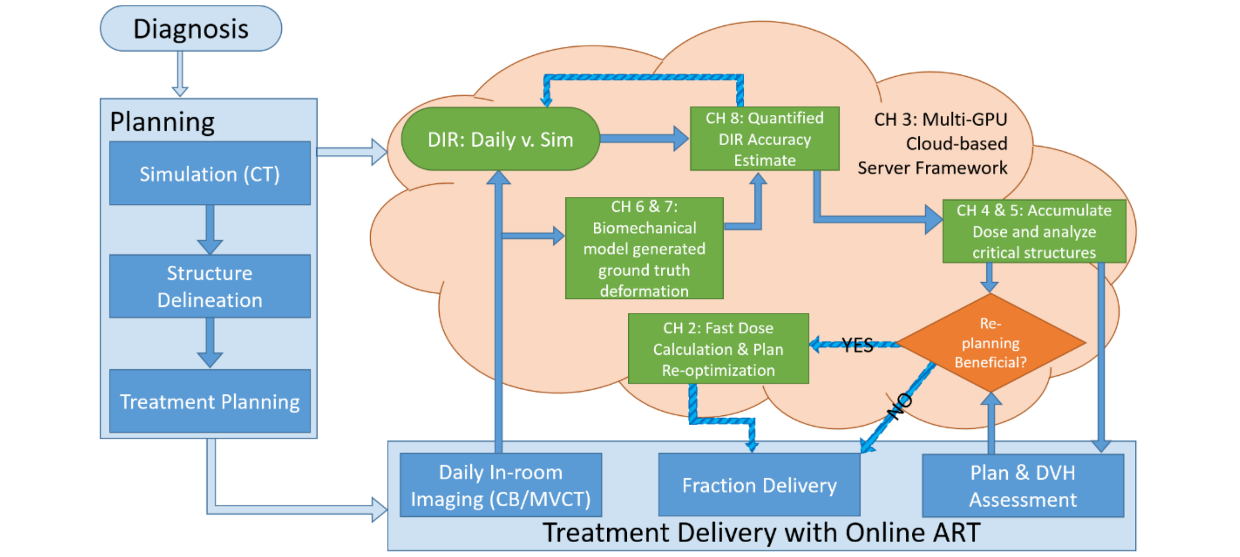

Diagnosis
Planning
Simulation (CT)
Structure Delineation
Treatment Planning
DIR: Daily v. Sim
CH 8: Quantified DIR Accuracy Estimate
CH 3: Multi-GPU Cloud-based Server Framework
CH 6 & 7: Biomechanical model generated ground truth deformation
CH 4 & 5: Accumulate Dose and analyze critical structures
CH 2: Fast Dose Calculation & Plan Re-optimization
YES
Re-planning Beneficial?
NO
Daily In-room Imaging (CB/MVCT)
Fraction Delivery
Plan & DVH Assessment
Treatment Delivery with Online ART

Figure 9.1. A potential workflow for on-line ART, combining the works of this dissertation.

computations remotely on a cloud-based server framework. The biomechanical model is already primed for a multi-GPU implementation, as each contoured structure can now be instantiated as an independent system of elements. With the pipeline established in chapter 3, porting GPU-based processes to the MGCS framework should just be a matter of intelligently dividing the data and tasks.

The projects were presented using head-and-neck applications, but they were developed to be versatile. Motion is a concern for several anatomical sites, such as lung, liver, and prostate. Specifically, the biomechanical model may be useful for sites such a liver, pancreas, and prostate, where soft tissue delineation can be difficult. The model can produce known deformations to fully characterize DIR performance within the visually homogenous soft tissue volumes.

Besides extending the current applications to other anatomical sites, biomechanical models have a wide variety of potential applications. One current area of development is elasticity estimations through inverse optimization analysis. They could also be used prospectively to project trends in weight loss and predict when intervention will become necessary. They could model internal motion to construct heterogeneous planning margins around the tumor, or to extrapolate volumetric deformations from two dimensional intra-fractional imaging, such as the images collected during ViewRay treatments. They could also be used to estimate internal deformations by registering the model surface with the patient surface observed during treatments by in-room 3D optical cameras.

The next logical step for the NVB dose algorithm is incorporation as the back-end dose calculation engine of a treatment planning system, which is currently being explored. Once implemented, a full validation study will need to be performed comparing the NVB dose output with commercial algorithms. I also believe that with the increased speed of the NVB implementation, some approximations intrinsic to the convolution/superposition algorithm, such as discretization of the energy spectrum and the density of spherical sampling during convolution, can be reduced or eliminated to more closely approach Monte Carlo dose simulations while sustaining the performance benefits of the convolution technique.

The next steps in the development for the DIR and dose accumulation framework should focus on user interface, robustness, producing concise, easy-to-interpret results, and testing the framework on additional imaging modalities. The published studies utilized weekly kVCTs, but further studies can be performed comparing MVCT and CBCT performance. Additionally, as the DIR algorithm is intensity-based optical flow, only slight modifications were required to analyze magnetic resonance (MR) imaging. A study is currently being pursued to analyze the anatomical variations over the treatment course using MR registrations and comparing the results with CT-based analysis. Ideally, the framework would be installed to run silently in the background of the clinic, analyzing data off-line and producing daily per-patient reports. This can be a fertile ground for collecting data, identifying trends, and improving the framework's tools. For instance, once NVB dose algorithm is validated within a TPS system, it can be incorporated into the DIR and dose accumulation framework for re-calculating dose on the daily anatomy, and comparing results to the dose warping analysis.

Lastly, the neural networks perhaps have the most potential impact on daily clinical operations. Neural networks are at the center of the current technological evolution, and are being applied to an incredibly wide-ranging number of fields. The work presented in chapter 8 only scratches the surface of possible applications for such work in radiotherapy. There is a wealth of annotated data in the form of past patients. The difficulty will be consolidating and sifting through that data with targeted and focused intent to construct reliable, informative, and accurate networks. Using GPUs for fast network training makes patient-specific networks feasible for future clinical use.

By utilizing the massive computational power of GPUs through intelligent parallelization and optimization, processes that are currently time-intensive can be made to run in real-time. Through the development of innovative tools like biomechanical modelling and neural networking, processes dependent on biased user-driven methods with small sample sizes can be automated and standardized. Combining these aspects with further improvements and integration, the tools presented here can facilitate on-line ART into the daily clinical workflow.