DEVELOPMENT AND VALIDATIONS OF A DEFORMABLE REGISTRATION ALGORITHM FOR MEDICAL IMAGES: APPLICATIONS TO BRAIN, BREAST AND PROSTATE STUDIES

Yangming Ou

A DISSERTATION

in

Bioengineering

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2012

Supervisor of Dissertation

___________________________
Christos Davatzikos, Ph.D.
Professor, Radiology

Graduate Group Chairperson

___________________________
Daniel Hammer, Ph.D., Professor, Bioengineering

Dissertation Committee

Andrew Maidment, Associate Professor, Radiology (Committee Chair)
Timothy Roberts, Professor, Radiology
Paul Yushkevich, Assistant Professor, Radiology
Nikos Paragios, Professor, Applied Mathematics, Ecole Centrale de Paris, France
DEVELOPMENT AND VALIDATIONS OF A DEFORMABLE REGISTRATION ALGORITHM FOR MEDICAL IMAGES:
APPLICATIONS TO BRAIN, BREAST AND PROSTATE STUDIES

© COPYRIGHT

2012

Yangming Ou

This work is licensed under the
Creative Commons Attribution NonCommercial-ShareAlike 3.0 License

To view a copy of this license, visit

http://creativecommons.org/licenses/by-nc-sa/3.0/
Dedicated to my family.
I would like to acknowledge the great help and support I have enjoyed from a long list of people, without whom this dissertation would not have been possible.

On the top of this list is my advisor, Dr. Christos Davatzikos. Dr. Davatzikos accepted my transfer from Virginia Tech in 2006, so that I got reunion with my then girl-friend and now wife. Over the years, he has carefully and patiently guided me into the beauties in science and engineering world – problem definition, literature review, critical thinking, experiment design, academic writing, scientific peer-review, argument presentation and defense, and more. Like many others, I have gone through dark and painful times in research and personal life; but Dr. Davatzikos’ support and understanding have been constantly there and a source of relief and motivation. I can name a lot of nice spirits I learned from him: persistence, patience, self-motivation, aiming ambitiously but proceeding conservatively. Yet the best I learned from him is the simplest rule: work hard even as smart as he is and even as successful as he has been.

My heartfelt thanks to the other four committee members: Drs. Andrew Maidment, Timothy Roberts, and Paul Yushkevich, here at UPenn, and Dr. Nikos Paragios, from Ecole Centrale de Paris, France. I am truly lucky to have them in my dissertation committee. Dr. Maidment has always reminded me the importance of critical thinking, exciting presentation and strong defense in ancient-style Greek philosophy, for I am pursuing a doctor in philosophy. Thanks also for his encouragement to broaden my scientific view, and to define clear paths for the future career. Dr. Tim Roberts, among the support of all other committee members, has helped reshape the course of this dissertation, and brought clinical meaning and impact into this work. His remark is a constant reminder of the fundamental questions that we, as hardcore engineers and computer scientists, have often times neglected: what can we do with science and technologies, what clinical relevance and significance can we
make. Dr. Paul Yushkevich, through his rich experience and personal success in the medical image analysis field, has kindly offered his support and his views on different research avenues in the field, therefore helped me refocus the interest on important, yet under-explored topics. I am fortunate to have his suggestions for academic writing and research directions. Dr. Nikos Paragios has been a close collaborator and a nice mentor whose guidance and supervision helped me fulfil many sections of this dissertation. Together with Dr. Davatzikos’ and Dr. Paragios’ support, my trip to Dr. Paragios’ lab in Ecole Centrale de Paris as an exchange student has been a memorable happy moment in my PhD life. The committee provides enormous help. Having them as mentors is a lifetime treasure.

Great appreciation also goes to three other professors that are important in different phases of my PhD study. I spent two years at the beginning of my PhD study at UPenn under the major supervision of Prof. Dinggang Shen, before he moved to a higher post at the University of North Carolina at Chapel Hill. Prof. Shen is a strict advisor with high academic standards and a nice friend with a caring heart. The basic medical image analysis knowledge I learned from him and the intensive training I received from him helped me fulfill many parts of this dissertation. Prof. Ragini Verma was supervising me in my first semester at UPenn. Over years she has offered tremendous help to my thinking, writing, and presentation. I am also truly impressed by her sharp and keen vision of how to apply the cutting edge technology into more clinically significant problems. Prof. Kilian Pohl came to UPenn in 2010. Through collaborations and communications, I am grateful for and really enjoyed his energy, motivation, courage and warm-hearted help.

UPenn provides a unique multi-disciplinary atmosphere gathering and attracting world-class clinicians and scientists. As a graduate student, I feel honored and cherish the opportunity to collaborate with faculties in different projects: Profs. Mitchell Schnall, Mark Rosen, John Tomaszewski, Michael Feldman, Mark Elliot for the prostate biopsy and histology-to-MRI registration projects; Prof. Nick Bryan for the diabetic study in the ACCORD project; Prof. Elias Melhem, Ronald Wolf, Drs. Seung-koo Lee, Sumei Wang for the multi-
parametric brain tumor detection project; Profs. Emily Conant, Susan Weinstein, Despina Kontos and Dr. Sarah Englander for the breast cancer longitudinal analysis project; and Prof. Michel Bilello for the brain tumor project.

The development of a stable, cross-platform, widely applicable, fully-automatic and publicly-released image piece of image registration software is a big project. Without the help and support from many people, DRAMMS software would never reach its current stage. In late 2008 and early 2009, I started with a few libraries and Gabor attribute code inherited from Drs. Yiqiang Zhan and Dinggang Shen in the lab. Through half a year of development, I finished the first fully-automatic version that features more than 20,000 lines of code. Accuracy was good but speed was slow. Luckily, Aristeidis Sotiras came from Prof. Nikos Paragios’ lab to visit our lab in the summer of 2009. During his visit he helped implement the discrete optimization based on Dr. Nikos Komodakis’ and Dr. Nikos Paragios’ FastPD package. This reduced the computation time from around 5 hours to around 30 minutes for a typical pair of brain images. This was the initial development phase and had led to Chapter 3 in this dissertation. Then I spent two years (2010 and 2011) running more than 400,000 registration jobs in large brain, cardiac, breast and prostate datasets, in comparison to more than 10 other registration software. The goal was to increase functionality of the software, and to optimize deformation mechanisms and the default parameters that are good for many different registration tasks on various organs. This was the software optimization and functionality enrichment phase and had led to Chapter 4. Another milestone through this process was when Andreas Schuh joined the lab. During mid-2011 to October 2012, Andreas Schuh and I worked closely to re-engineered the software, making it cross-platform (Linux, Mac), nifti-compatible, and much easier and more powerful to use. A typical schedule of our workday was like this: I would leave the lab at around 7pm and Andreas left the lab around 10pm; we continued to work at home and to exchange emails until 1-2am. When I got up the next morning 7-8am, I usually found an email from Andreas sent from 3-4am happily telling me a bug had been resolved. Then I would go on with experiments and coding, emailing him with more bugs in the morning, waiting for his arrival at around 11am;
then we work together and repeat another day. To me, Andreas’ working style is a vivid explanation of the delicate German engineering. Besides developers, I deeply appreciate all the constructive feedback in the 8 major rounds of tests throughout the last 2 years from the following careful testers: Jimit Doshi, Xiao Da, Harsha Battapady, Guray Erus, Alex Smith, Dong Hye Ye, Madlura Ingalhalikar, Harini Eavani, Dongjin Kwon, Drew Parker, Erdem Varol, Hamed Akbari, Bilwaj Gaonkar, Vanessa Clark, Steffen Wachenfeld here at UPenn, and Nikolaos Koutsouleris (Ludwig-Maximilians-University, Germany), Helene Langet (GE Healthcare/Ecole Centrale de Paris, France), Lena Rademacher (The University Hospital Aachen, Germany) and Marcus Zanetti (University of San Paulo, Brazil).

My appreciation also goes to my fellow lab members. The happy time and the friendship with this group of smart people is a life treasure for me. Some of them have already been mentioned above as DRAMMS developers and testers. Others include: Sajjad Baloch, Kayhan Batmanghelich, Elena Bernardis, Priyanka Bhatt, Luke Bloy, Hongmin Cai, Houwei Cao, Rong Chen, Aoyan Dong, Yong Fan, Chaolu Feng, Roman Filipovych, JingJing Gao, Yasser Ghanbari, Ali Gooya, Jihun Hamm, Stathis Kanterakis, Parmeshwar Khurd, Zhiqiang Lao, Chunming Li, Naixiang Lian, Arvind Rao, Hari Sundar, Peng Wang, Ying Wang, Guorong Wu, Xiaoying Wu, Feng Xu, Zhong Xue, Jinzhong Yang, Rachel Yotter, Evangelia Zacharaki, Tianming Zhan, Yiqiang Zhan, Tianhao Zhang, and Yuanjie Zheng.

Thanks also to my friends in Ecole Centrale de Paris, for being part of a memorable journey for me in France: Pierre-Yves Baudin, Ahmed Besbes, Salma Essafi, Nicolas Honnorat, Fabrice Michel, Sarah Parisot, Loic Simon, Olivier Teboul, Chaohui Wang, and Bo Xiang.

Special thanks to the following nice ladies and gentleman in administrative roles, for the help and support: Paraskevi Parmpi, Amanda Shacklett and Mark Bergman at SBIA lab, Kathleen Venit in the Department of Bioengineering, and Betty Gentner in the engineering school.
The revision of this dissertation is under the great help from Amanda Shacklett, Guray Erus, Alex Smith, Aristeidis Sotiras, Elena Bernardis and Bilwaj Gaonkar. Thank you for helping polish this documentation.

My utmost gratitude goes to my family, to whom this dissertation is dedicated, for their endless and unconditional love and for my deepest love in them. PhD is a long journey. For me and for my family it has special meanings. The first two years of my PhD were unfortunate: three grandmas and grandpas passed away, two of them within 10 days. The last two and a half years were filled with a lot of pressure. My graduation had to be postponed a couple of times. My wife had to relocate for work, and later went through a serious illness that almost left her with life-long after-effects. My mother was diagnosed with early-stage cancer and underwent a surgery. Soon after this my father was suddenly hospitalized months before his 60th birthday, and was diagnosed with end-stage metastatic colon cancer. Since then, he has been fighting with death every day. For this strike and the overwhelming sadness, my mother went blind in one eye. As a student and a young professional, I concentrate on the software release, PhD work and other lab projects. As a husband, I drove or flew frequently for emotional support, but felt sorry for being unable to financially support. As a son and the only child, I took parents to Philadelphia, taking care of them, cherishing every minute with Dad, and accompanying him for weekly chemotherapy. Seeking treatment here in US is especially an almost impossible task for a poor foreigner. For my love in him, I had tirelessly and fearlessly fought with numerous medical, financial, legal and immigration issues. From the desert of despair, fortunately there is glimmering oasis of hope. Life has changed my family. Life has changed me. Out of tears, sorrow and even despair, life teaches me what love is, life teaches me to calm, to be brave, to be always optimistic, and to love. Today, death is still constantly waving for my father. However, with my wife having finally moved back to the east coast, my mother having survived from early-stage cancer, my father miraculously still alive at this very minute, the finish line for the Ph.D. degree eventually within reach, and a new life finally on the way, I am thankful and hopeful.
ABSTRACT

DEVELOPMENT AND VALIDATIONS OF A DEFORMABLE REGISTRATION ALGORITHM FOR MEDICAL IMAGES:
APPLICATIONS TO BRAIN, BREAST AND PROSTATE STUDIES

Yangming Ou

Christos Davatzikos, Ph.D.

This dissertation presents work on deformable registration of medical images. Deformable registration is a fundamental problem in medical image computing, and is central in analytic methods for understanding population trends of imaging phenotypes, for measuring longitudinal change, for fusing multi-modality information, and for capturing structure-function correlations.

Despite over 20 years of extensive research and technology innovations, two limitations still exist in the literature of image registration, which are central in this dissertation. The first one is the ambiguity in determining anatomical correspondences between images. This is caused by the fact that most registration methods match images based on image intensities. However, intensity alone does not necessarily represent the anatomical or geometric context in the images. The second challenge is the problem of missing correspondences. This arises from the presence of pathologies in images, whose correspondences are not present in the images of healthy subjects. Moreover, many registration methods are fine-tuned to a particular problem, thereby losing generality. The proposed work contributes towards overcoming these limitations. It reduces ambiguity by matching voxels based on their geometric contexts instead of intensities alone. It handles missing correspondences by a newly developed mutual-saliency metric, which automatically identifies the regions having difficulty finding correspondences in the other image, and reduces their negative impact. The proposed method is designed in a general way in that it does not rely on any task-
specific annotations of tissue, structure or feature points.

Extensive experiments are presented to demonstrate generality, accuracy and robustness, which are the key properties of the proposed method. Experiments involve images of various organs (brain, breast, heart) in various registration settings (cross-subject, longitudinal) on various public and in-house datasets, and include quantitative comparisons with many state-of-the-art methods.

To demonstrate the wide application of the proposed method in clinical and research studies, five topics utilizing the proposed method are presented. They are examples of population studies, longitudinal studies, and atlas-based segmentations.

The proposed method is fully-implemented, extensively-tested and publicly-released to meet the growing needs of many large-scale clinical and academic studies.
# TABLE OF CONTENTS

**ACKNOWLEDGEMENT** ................................................................. iv

**ABSTRACT** ................................................................................... x

**LIST OF TABLES** ................................................................. xiv

**LIST OF FIGURES** ............................................................... xv

**CHAPTER 1: Introduction** ................................................. 1

1.1 Background and Motivations .................................................. 1

1.2 Dissertation Overview ........................................................... 11

1.3 Thesis Organization ................................................................. 15

**CHAPTER 2: Literature of Image Registration** .................. 16

2.1 Image Registration: General Formulation ............................. 17

2.2 Classification of Existing Registration Methods .................. 18

2.3 Landmark-based Registration Methods ............................... 19

2.4 Voxel-wise Registration Methods ........................................... 23

2.5 Some Publicly-Available Image Registration Tools ............. 32

**CHAPTER 3: DRAMMS Algorithm** ...................................... 42

3.1 Technical Motivations ............................................................. 43

3.2 DRAMMS Methods ................................................................. 46

3.3 Numerical Implementation ..................................................... 65

3.4 Results ................................................................................... 69

3.5 Discussion .............................................................................. 78

**CHAPTER 4: Validations of DRAMMS** ............................... 85
4.1 In Registering Skull-stripped Brain MR Images Across Subjects . . . . . . . 86
4.2 In Registering Raw Brain MR Images Across Subjects . . . . . . . . . . . . 103
4.3 In Registering Lesion-bearing Brain to Healthy Brain MR Images . . . . . 116
4.4 In Registering Tumor-Recurring Brain to Healthy Brain MR Images . . . . 122
4.5 In Registering Cardiac MR Images Across Subjects . . . . . . . . . . . . . . 138
4.6 In Registering Breast Cancer Images Over Time . . . . . . . . . . . . . . 150
4.7 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 160

CHAPTER 5 : Applications of DRAMMS . . . . . . . . . . . . . . . . . . . . . . 162

5.1 Construction of Statistical Atlases for Brain Lesions in Diabetic Population 163
5.2 Generation of Post-Surgery Brain Tumor Recurrence Maps . . . . . . . . . 170
5.3 Quantification of Longitudinal Breast Cancer Changes . . . . . . . . . . . . 173
5.4 Multi-Atlas-based Skull-stripping for Brain Images . . . . . . . . . . . . . . 177
5.5 Multi-Atlas-based Segmentation of the Prostate . . . . . . . . . . . . . . . 185
5.6 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 193

CHAPTER 6 : Towards Improving DRAMMS Speed and Accuracy . . . . . . . . . 195

6.1 discreteDRAMMS: Landmark Detection and Matching . . . . . . . . . . . . 196
6.2 Tissue Segmentation Improves Brain Registration Accuracy? . . . . . . . . 207

CHAPTER 7 : Discussion and Conclusions . . . . . . . . . . . . . . . . . . . . . . 218

7.1 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 218
7.2 The DRAMMS Algorithm . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 220
7.3 Validations of DRAMMS . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 223
7.4 Applications of DRAMMS . . . . . . . . . . . . . . . . . . . . . . . . . . . . 225
7.5 Towards Improving DRAMMS Speed and Accuracy . . . . . . . . . . . . . 227
7.6 Concluding Remarks . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 229

APPENDIX . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 231

CHAPTER A : DRAMMS Software Package . . . . . . . . . . . . . . . . . . . . . 231
LIST OF TABLES

TABLE 1: Comparing the computational times between the discrete optimization and gradient descent optimization. 67

TABLE 2: A List of how DRAMMS bridges the gap between voxel-wise and landmark/feature-based registration methods. 79

TABLE 3: Summary of the NIREP and LONI-LPBA40 datasets. 88

TABLE 4: Summary of the registration methods compared in this study. 91

TABLE 5: Lists of patients in the responder and non-responder groups. 153

TABLE 6: Comparison of our skull-stripping results to ROBEX results in the same datasets. 181

TABLE 7: Deformation smoothness before and after landmark clustering. 204

TABLE 8: Spatial dispersion of landmarks before and after landmark clustering. 205

TABLE 9: Statistical significance of the improved regional overlap. 216
LIST OF FIGURES

FIGURE 1 : Visualization of the image registration concept. . . . . . . . . . . . 2
FIGURE 2 : Motivational Study 1. . . . . . . . . . . . . . . . . . . . . . . . . . 4
FIGURE 3 : Motivational Study 2. . . . . . . . . . . . . . . . . . . . . . . . . . 5
FIGURE 4 : Motivational Study 3. . . . . . . . . . . . . . . . . . . . . . . . . . 7
FIGURE 5 : Motivational Study 4. . . . . . . . . . . . . . . . . . . . . . . . . . 8
FIGURE 6 : Motivational Study 5. . . . . . . . . . . . . . . . . . . . . . . . . . 10
FIGURE 7 : Sketch of the image registration process. . . . . . . . . . . . . . . 17
FIGURE 8 : Free form deformation (FFD) model in 2D. . . . . . . . . . . . . . 29
FIGURE 9 : Need for mutual-saliency weighting: pathological images. . . . . . 44
FIGURE 10 : Need for mutual-saliency weighting: normal images. . . . . . . . . 45
FIGURE 11 : DRAMMS framework. . . . . . . . . . . . . . . . . . . . . . . . . 47
FIGURE 12 : Example Gabor attribute images. . . . . . . . . . . . . . . . . . . 50
FIGURE 13 : The role of full Gabor attributes. . . . . . . . . . . . . . . . . . . 53
FIGURE 14 : Sketch of the selection of training voxel pairs. . . . . . . . . . . . 56
FIGURE 15 : Optimization of the cost function for attribute selection. . . . . . 58
FIGURE 16 : The role of Gabor attributes. . . . . . . . . . . . . . . . . . . . . 59
FIGURE 17 : The idea of the mutual-saliency measure. . . . . . . . . . . . . . . 62
FIGURE 18 : Sketch of the mutual-saliency definition. . . . . . . . . . . . . . . 62
FIGURE 19 : The role of mutual-saliency maps. . . . . . . . . . . . . . . . . . . 64
FIGURE 20 : Image similarity after registration. . . . . . . . . . . . . . . . . . 71
FIGURE 21 : Brain images for brain atlas construction. . . . . . . . . . . . . . . 72
FIGURE 22 : The constructed brain atlases. . . . . . . . . . . . . . . . . . . . . 73
FIGURE 23 : Comparing registration methods in brain atlas construction. . . . 74
FIGURE 24 : Comparing accuracies of registration methods in each subject. . . 75
FIGURE 25: Registration results between the histological and MR images of the same prostate. 77

FIGURE 26: The importance of using the proper parameters for attribute extraction. 80

FIGURE 27: Typical subjects in the NIREP and LONI-LPBA40 datasets. 89

FIGURE 28: Registration of 3D brain MR images from two different NIREP subjects by DRAMMS. 95

FIGURE 29: Registration accuracies in the NIREP and LONI-LPBA40 datasets. 97

FIGURE 30: Surrogates for the aggressiveness of deformations (the NIREP dataset). 98

FIGURE 31: Surrogates for the aggressiveness of deformations (the LONI-LPBA40 dataset). 99

FIGURE 32: Deformation accuracy-aggressiveness correlation. 101

FIGURE 33: Three typical subjects from the BrainWeb dataset. 106

FIGURE 34: Typical raw brain images from the IBSR, OASIS and ADNI datasets. 108

FIGURE 35: Sketch of the definition of distance error between two regions. 110

FIGURE 36: Registration accuracy in the BrainWeb dataset. 111

FIGURE 37: Typical DRAMMS registration results in within-dataset registrations. 112

FIGURE 38: Within-dataset registration accuracy. 113

FIGURE 39: Cross-dataset registration accuracy. 114

FIGURE 40: Registration accuracy on brain masks for all 150 cross-dataset registrations. 114

FIGURE 41: Brain images with simulated lesions. 118

FIGURE 42: Deformation errors caused by the presence of lesions. 121

FIGURE 43: Measure registration accuracies in different zones. 126

FIGURE 44: Registration of a brain image with tumor recurrence to a normal brain template by DRAMMS. 129
FIGURE 66: Dice scores for different numbers of templates (atlases) ... 183
FIGURE 67: The pipeline for segmentation of the prostate. ... 187
FIGURE 68: Prostate segmentation accuracy in 6 typical training subjects... 191
FIGURE 69: Visualization of segmentation results in 6 randomly chosen subjects from training set. ... 192
FIGURE 70: The three steps in discreteDRAMMS. ... 198
FIGURE 71: Displacement errors compared to ground-truth deformation. ... 204
FIGURE 72: Landmark correspondences established in real images containing tumor-induced missing correspondences. ... 206
FIGURE 73: Overview of the experiment design for studying the effect of automatic tissue segmentation to registration accuracy. ... 210
FIGURE 74: Comparing registration accuracies between using intensity and using tissue segmentation for registration. ... 215
FIGURE 75: Homepage of the DRAMMS website. ... 232
FIGURE 76: Configuration of cmake. ... 237
FIGURE 77: An example of how to use dramms main script to calculate RAVENS maps. ... 244
FIGURE 78: The 12 example registration tasks included in the tutorials. ... 247
FIGURE 79: Registration of two 2D simulated images. ... 248
FIGURE 80: Registration of two 2D multi-modality (histology and MR) slices of a mouse brain. ... 248
FIGURE 81: Registration of 3D brain MR images from two different subjects (skull-stripped) in NIREP dataset. ... 250
FIGURE 82: Typical DRAMMS registration results in within-dataset registrations ... 252
FIGURE 83: Typical DRAMMS registration results in across-dataset registrations ... 252
FIGURE 84: Longitudinal registration of pediatric images. ... 255
FIGURE 85: Registration results between lesion-bearing and normal images ... 257
FIGURE 86: DRAMMS results between tumor-recurring and normal brain images.

FIGURE 87: The same set of results as in Fig. 86, but in coronal view.

FIGURE 88: Registration of longitudinal mouse brain images to follow mouse brain development.

FIGURE 89: DRAMMS registration results for inter-subject cardiac images.

FIGURE 90: Registration of longitudinal cardiac images of the same subject.

FIGURE 91: DRAMMS results on longitudinal breast cancer images.

FIGURE 92: Registration of multi-modality images of a same prostate.
CHAPTER 1

Introduction

1.1. Background and Motivations

Thanks to fast-evolving medical imaging technologies, medical images are now indispensable tools in routine clinical practice and many biomedical studies. Medical images, such as magnetic resonance imaging (MRI), computerized tomography (CT) and X-rays, reveal anatomical structures that assist doctors in diagnosing abnormalities. Medical images, such as functional MRI (fMRI), positron emission tomography (PET), and single photon emission tomography (SPECT), offer functional information that increases our understanding of the relationship between structures and functions. Medical images of the same patient taken at a series of time points offer an invaluable tool for tracking disease progression and monitoring treatment effects. Medical images acquired before surgery aid in treatment planning and provide guidance for the surgery. Medical images from diseased populations provide opportunities to understand disease-related imaging bio-markers. In recent years, new imaging technologies allow visualizations of structures and pathologies in even microscopic and molecular resolutions. Today, researchers and doctors are equipped with a wide spectrum of images: at various resolutions, from a series of time points, and from multiple imaging modalities.

The explosion of medical imaging data calls for highly automated and sophisticated image
analysis technologies. This dissertation contributes to one of the fundamental image analysis technologies, known as "medical image registration", or simply "image registration". Image registration is the process of transforming different images into a common spatial coordinate system. That is, after registration, the same spatial location in different images correspond to the same anatomical structure (Zitova and Flusser, 2003; Crum et al., 2004). Fig. 1 visualizes the concept of image registration. Given the source and the target images, image registration computes a transformation (Fig. 1(d)) that transforms the source image (Fig. 1(a)) into the space of the target image (Fig. 1(b)). The transformation is also named deformation in deformable image registration. The result of image registration is the registered image (Fig. 1(c)) that is in anatomical correspondence with the target image (Fig. 1(b)) at every spatial location. It is also said that the registered image resides in the same coordinate system as the target image. The computed deformation records how the source image deforms into the target image space for each and every voxel location in the source image. Vital to the computation of the deformation, hence vital to image registration, is to accurately establish correspondences between the two images at the voxel level.

Image registration has been at the forefront of medical image analysis research for the past two decades. It is the building block for many clinical and research studies. In population studies, for instance, image registration brings the images from a population of subjects into the same coordinate system. This enables the joint analysis of multiple subjects to reveal disease-related imaging phenotypes. In longitudinal studies, image registration captures
how the images deform over time. This leads to the quantification of disease changes, allowing us to observe disease progression and to evaluate treatment effects. In image-guided surgeries, image registration links the offline, high-resolution pre-operative images and the real-time, low-resolution intra-operative images, to provide surgical guidance. In computer-assisted diagnosis (CAD), such as the detection of brain tumors or the screening of prostate or breast cancers, image registration brings together multiple images obtained from various imaging modalities (e.g., CT, MRI, PET, etc). This enables the fusion of complementary information from multi-modality data to more accurately identify abnormalities at early disease stages.

To satisfy the need for all the aforementioned studies, it is desirable to have an image registration method that is general and accurate. Being general means that the same image registration method should be readily applied to images containing various organs (brain, breast, heart, prostate, etc), and should be readily applied to various image registration tasks (cross-subject, longitudinal, multi-modality, etc). In the past two decades, many image registration methods have been developed. However, they are either task-specific, targeting particular problems, or they are general, but often of relatively limited accuracy. Generality is difficult to achieve, because different registration tasks pose different challenges. For example, the challenge in cross-subject brain registration mainly stems from large anatomical differences among different individuals, whereas the challenge in longitudinal breast registration is the presence of large deformations, because the breasts contain soft tissue that is more deformable than brain tissues. On the other hand, accuracy for a general method is also a high requirement. Accuracy is especially difficult to achieve when there is significant anatomical differences (such as when pathologies are present), or when there are significant image appearance differences (such as in multi-site images, where different imaging protocols bring about greatly different image contrasts, different image resolutions, and even different image fields-of-view). In such situations, an image registration method should be robust with regard to anatomical and appearance differences to gain high registration accuracy.
Figure 2: Motivation 1: the need for a robust registration method to account for the missing correspondences between lesion-bearing and normal brain images. Red contours contain lesions whose correspondences are not present in the normal template.

To demonstrate the challenges mentioned above, below are five typical research and clinical studies where image registration is needed. They are examples of population studies, longitudinal studies and atlas-based segmentation studies. They are in particular the motivations for developing a new image registration algorithm in this dissertation.

Study 1. Tracking Treatment Effects in Diabetic Population

Type II diabetes patients are often associated with white matter lesions (WMLs) in the brain (Schmidt et al., 2004; Jongen et al., 2007; Anan et al., 2009). MRI is often used to observe WMLs due to its high sensitivity and its high contrast in the deep brain structures (Anan et al., 2009; Starr et al., 2003). In the Action to Control Cardiovascular Risk in Diabetes (ACCORD) clinical trial (Buse et al., 2007), one aim is to study the change of the average WML load in Type II diabetic population as the result of different treatment strategies. To obtain the average WML loads in a population, a major step is to register,
Figure 3: Motivation 2: the need for a robust registration method to account for tumor-induced missing correspondences between images having recurrent tumors and the healthy template image, which serves as anatomical reference. Red arrows point out recurred tumors.

or spatially normalize, the lesion-bearing brain MR images from all diabetic patients into a common template image space where they can be averaged and compared (Fig. 2). The template image is usually chosen from a healthy subject outside this diabetic population, to better represent the normal anatomy. Registering each patient’s lesion-bearing brain image to the healthy template image, however, poses considerable challenges to most existing image registration methods, because of the missing correspondences (lesions being present in the patient image but not in the healthy template image). Therefore, this motivates the development of a new image registration algorithm that is robust to the lesion-induced missing correspondences in cross-subject registration tasks.

**Study 2. Studying the Spatial Distribution of Recurrent Brain Tumors**

Brain tumors are intracranial solid neoplasms that often pose serious life threats. The
standard treatments for brain tumors include surgical resection, followed by chemotherapy and/or radiotherapy. Because of the highly invasive and infiltrative nature, brain tumors often recur after treatment. Some questions of interest are: what is the relationship between the spatial locations of the recurrent tumor and the spatial locations of the original tumor; and does tumor recurrences in different patients follow certain common patterns (for example, do tumors recur along major vessels, white matter pathways, or stem-cell generating brain regions in most patients)? One straightforward approach to answer the questions above is to register the tumor-bearing brain images (often MRI) from a large number of patients into a common template space that expresses normal brain anatomy and serves as spatial reference (Fig. 3). We can thus generate statistical maps of the spatial distribution of tumor recurrences, which can help us address the aforementioned questions. Similar to Study 1, the registration between the tumor-bearing brain images to the healthy template image also encounters the missing correspondences problem – recurrent tumors being present in patient images but not in the healthy template image. The missing correspondence problem here is even more challenging than that in Study 1, as tumors usually occupy larger space than the white matter lesions and often don’t have a clear boundary because of the infiltrative nature. Therefore, this study motivates the development of a new image registration algorithm that can effectively, and automatically (i.e., without expert-segmentation of tumors) handle the tumor-induced missing correspondences.

**Study 3. Quantifying Longitudinal Cancer Changes in Breast Cancer Patients**

According to the American Cancer Society, breast cancer is the second most frequent cancer after skin cancer, and the second leading killer after lung cancer in US women [American Cancer Society, 2012]. Patients with breast cancer often receive chemotherapy treatment regime that spans weeks to months. During the course of chemotherapy, patients are often followed with periodic breast MRI scans. The longitudinal images, as shown in Fig. 4 allow us to observe cancer changes and to assess treatment effects. In the current clinical paradigm, doctors measure the volumes of breast tumors in longitudinal images and
Figure 4: Motivation 3: the need for a new registration algorithm to effectively capture the longitudinal cancer change in breast cancer patients receiving chemotherapy. Longitudinal registration of breast cancer images poses different challenges from the ones described earlier for brain images. Please refer to the main text for more detail.

report the volumetric change over time. However, doctors’ measurement is subjective, time-consuming and of low reproducibility. More importantly, it only conveys the change of the entire breast tumor. It does not allow us to quantify the spatial inhomogeneity within breast tumors, which might help us understand how different sub-regions of breast tumors respond to the treatment differently. To address these issues, it is desirable to use automated image registration techniques to quantify the longitudinal breast tumor change at the voxel level, which allows us to fully characterize the tumor change morphologically. Based on such quantifications, we can further test: 1) whether those patients who respond well to chemotherapy (i.e., responders) share a common cancer change patterns (e.g., rate, locations), and 2) whether, and if so, how, responders and non-responders differ in their cancer change patterns. Perhaps more importantly, we can relate image voxels in the follow-up scans to voxels in the baseline scans, and therefore relate baseline imaging characteristics to treatment response, in cases of tumors with heterogeneous imaging profiles.

However, registration of longitudinal breast images is not trivial. Compared to the brain or the heart, the breasts contain soft tissue that is highly deformable. This is further complicated by the positioning difference of the same breasts at different times, by the inclusion of chest structures in the breast images, and by the large change of breast tumors in shape, size and texture. Therefore, this study motivates the development of a new image
registration method that 1) applies to human organs other than the brain (e.g., the breast); 2) accurately identifies anatomical correspondences over time; and 3) is robust with regard to the background noise and to the confounding effects of surrounding chest structures.

**Study 4. Extracting the Brain in Raw Brain Images (Skull-stripping)**

In many medical image analysis tasks involving brain images, the first step is to remove the background noise and extracranial tissues, and to focus on the brain only. This step is often known as brain extraction, or skull-stripping. Skull-stripping is an important step, because any errors in skull-stripping would propagate to all analysis steps that follow. Therefore, the requirement for accuracy and robustness in the skull-stripping step is high. Traditionally, skull-stripping is done either manually by some human experts or automatically by some computerized methods [Smith 2002; Kovacevic et al. 2002; Segonne et al. 2004; Zhuang et al. 2006]. Manual skull-stripping is time-consuming and of low reproducibility. Automatic approaches are fast and objective, but are often subject to considerable errors.
Among many automatic skull-stripping approaches developed in the literature, multi-atlas methods have shown great promise and have attracted growing attentions (e.g., Lötjönen et al., 2010). The idea is to register multiple brain images (known as atlases), whose brains have already been labeled (often by human experts), onto the target image. Registration enables us to pass on, or propagate, the brain labels from the atlases into the target image space, as shown in Fig. 5. In this multi-atlas skull-stripping framework, image registration directly affects the accuracy in propagating brain labels from atlases to the target space. However, it is not rare to observe low accuracy or even failures using many of the existing image registration methods. This is due to the large anatomical variations, the high levels of noise and field inhomogeneity found in routine clinical images, and the different fields-of-view (FOVs). The failure rate increases when dealing with multi-site datasets, which are the norms in large-scale studies. One example multi-site dataset is in the Alzheimer’s Disease Neuroimaging Initiative (ADNI) study (http://adni.loni.ucla.edu/). The ADNI dataset includes images from more than 800 normal subjects, Mild Cognitive Impairment (MCI) patients, and Alzheimer’s Disease (AD) patients. The images are acquired in multiple institutions to investigate AD-related imaging, biochemical and genetic bio-makers. They present notable differences in some imaging characteristics, such as field of view. As a result, some registration methods may fail in as many as one-fifth or one-fourth of the cross-subject registration tasks. This motivates the development of a new image registration method that is robust and accurate in cross-subject registration of raw brain images acquired with relatively more variable imaging characteristics.

Study 5. Segmenting the Prostate in Prostate MR Images

Prostate cancer is the most common cancer and the second leading cause for cancer-related death in American men, with an estimate of 241,740 new cases and 28,170 deaths in the year 2012 (American Cancer Society, 2012). Accurate segmentation, or localization, of the prostate is essential for the effective planning of prostate brachytherapy (Pasquier et al., 2007; Gao et al., 2010). With the accurate segmentation of the prostate, doctors can
concentrate the radiation dose on the prostate, and spare the surrounding structures such as the bladder and the rectum. There are many approaches to segment the prostate from prostate images. Similar to the brain extraction in Study 4, multi-atlas-based segmentation of the prostate has shown promise in recent studies (e.g. [Klein et al., 2008]). As Fig. 6 shows, image registration can transfer the known prostate segmentation masks from multiple atlases into the target image space. The transferred prostate masks are then fused into one final prostate segmentation mask in the target image space. The accuracy in prostate segmentation is mainly dependent on the accuracy in registering prostate atlases into the target image space. Registration of prostate images is usually more difficult than registration of raw brain images, because of imaging artifacts, inter-subject anatomical variabilities, differences in rectum and bladder structures (the rectum and the bladder dominate the image space), and the associated FOV differences. Because of the reasons above, many existing registration methods may fail in the prostate image registration task. This also
motivates the development of a new image registration algorithm that achieves robustness to all the aforementioned confounding factors.

**Summary of Motivations**

The aforementioned five representative studies emphasize the need for a new medical image registration method, which should satisfy the following requirements:

1. Fully-automatic, with no need for human interventions and no need for prior knowledge (e.g., tissue/structure segmentations);

2. General, in the sense that i) it should be generally applicable to images of various organs (brain, breast, prostate, etc) without heavy case-specific parameter tuning; and ii) it should be generally applicable to different registration settings (cross-subject registration, longitudinal registration, etc);

3. Accurate, in establishing anatomical correspondences between images;

4. Robust, with regard to the background noise, intensity inhomogeneity, anatomical variation, and even missing correspondences caused by pathologies or by FOV differences.

Developing an automated medical image registration method that satisfies the desirable generality, accuracy and robustness properties is therefore the overall goal of this dissertation.

**1.2. Dissertation Overview**

This dissertation develops a new medical image registration algorithm, validates its generality, accuracy and robustness in brain, cardiac and breast images, and shows its wide applications as represented in the five studies mentioned previously in this chapter. The last part of this dissertation describes approaches that aim to further improve the speed
and accuracy of the proposed image registration method. Also, the proposed method is now a fully-implemented, extensively-tested, and publicly-available software (http://www.rad.upenn.edu/sbia/software/dramms/), which is briefly introduced in the Appendix to meet the demands in large clinical and academic studies.

The proposed image registration method is termed "DRAMMS", short for "Deformable Registration via Attribute Matching and Mutual-Saliency Weighting". The following subsections will briefly introduce the five major parts of this dissertation: the DRAMMS algorithm, validations of DRAMMS, example applications of DRAMMS in research and clinical studies, extensions of DRAMMS, and the DRAMMS software.

**Overview of the DRAMMS Algorithm and Its Contributions**

DRAMMS deals with two fundamental challenges in image registration: 1) how to accurately establish voxel-wise correspondences; and 2) how to effectively and automatically handle missing correspondences, or outlier structures that exist in one image but not the other. To deal with the first challenge, DRAMMS matches image voxels by their geometric contexts. Compared with the image intensity information that is used in most existing image registration methods, the geometric context information as used in the DRAMMS framework renders each voxel more distinctive for finding correspondences. Approaches to extract the geometric attributes and to select the optimal components are developed. This is the first major contribution of DRAMMS. To deal with the challenges caused by missing correspondences, DRAMMS automatically identifies regions that have difficulties establishing correspondences between images, and reduces their negative impact during registration. This is handled by a newly-developed metric named "mutual-saliency". The development of the mutual-saliency metric and the use of it to modulate image registration is the second contribution in the DRAMMS algorithm.

**Overview of Validations of the Proposed DRAMMS Algorithm**

The DRAMMS algorithm is extensively validated for its general applicability, robustness
and accuracy. To demonstrate its general applicability, DRAMMS is tested in various image registration tasks (cross-subject, longitudinal) involving images of various organs (brain, heart, breast). To demonstrate its robustness, DRAMMS is tested in registration experiments containing very large anatomical variations and missing correspondences (caused by pathologies or FOV differences). To demonstrate its accuracy, DRAMMS is qualitatively and quantitatively compared with up to 12 general-purpose methods on public and in-house datasets. At the end of the validation chapter, we also report registration tasks and scenarios where DRAMMS does not apply or may fail.

Overview of Example Applications of the Proposed DRAMMS Algorithm

The dissertation then shows the wide application of the validated DRAMMS method in various clinical and research studies. Five studies are presented. They are the ones mentioned previously in this dissertation as motivations for the development of the DRAMMS algorithm. They are examples of population studies, longitudinal studies, and atlas-based anatomy annotations. When applied to registering lesion-bearing brain images onto the normal brain template, DRAMMS enables us to derive the statistical maps of white matter lesion loads, from which the spatial distributions of lesions can be studied. Based on that, it is possible to examine the effects of different treatment strategies. When applied to registering brain images with tumor recurrence to the normal brain template, DRAMMS transfers the recurrent tumor regions from images of a number of patients into a same template space that represents the healthy anatomy. This offers an opportunity to observe tumor recurrence patterns in brain tumor populations, and to relate them to clinical, demographic, genetic and other variables. When applied to registering longitudinal breast cancer images, DRAMMS helps quantify, at the voxel level, the volumetric change of breast tumors over time. Based on that, we can test the hypothesis that the patients responding well to chemotherapy and the patients not responding well to chemotherapy exhibit different patterns in how their breast tumors change. When applied to registering raw brain or prostate MR images, DRAMMS enables us to transfer the expert-defined brain or prostate
segmentation masks from multiple atlases to the target image. After several post-processing steps (e.g., label fusion), this results in the extraction of the brain or the prostate in the target image, with an accuracy comparable to that of the human expert’s delineations.

**Overview of Potential Improvements of the Proposed DRAMMS Algorithm**

In the last part of this dissertation, two attempts are made toward the further improvement of the speed and accuracy of DRAMMS. Towards reducing the computational time of DRAMMS, attempts are explored to discretize DRAMMS — instead of establishing correspondences in every image voxel, we can establish correspondences at a subset of voxels that are automatically identified. Correspondences found on the subset of voxels preserve the accuracy of the DRAMMS algorithm. They can then be used to guide other intensity-based registration methods that are computationally efficient. Towards higher accuracy, one general approach is to incorporate the task-specific prior knowledge into the DRAMMS framework. The prior knowledge can be expert-defined structural annotations, or automatic tissue segmentations. The latter is used, and its potential to improve the accuracy of image registration is tested in cross-subject brain registration tasks.

**Overview of the DRAMMS Software**

Since dissemination of our work is important in view of the relative maturity of the image registration field, and also because of the needs faced by many large clinical studies, DRAMMS is fully implemented, extensively tested, and publicly released as a command-line software running in the UNIX/Linux or Mac operating systems ([http://www.rad.upenn.edu/sbia/software/dramms/](http://www.rad.upenn.edu/sbia/software/dramms/)). Software manuals and Tutorials are provided in the Appendix of this dissertation. Parameters can be default, which leads to reasonable results in many cases, or can be tuned as several tutorial examples suggest. The DRAMMS software also provides a rich set of auxiliary tools for the operations of the obtained deformations. The operations include using the obtained deformation to warp another image, calculating the Jacobian determinant maps which quantify the volumetric change at the voxel level, calculating the RAVENS tissue density maps, and invert-
ing/smoothing/composing/adding/subtracting/averaging one or two deformations.

1.3. Thesis Organization

The rest of this dissertation is organized as follows.

Chapter 2 reviews the literature of medical image registration.

Chapter 3 identifies two fundamental challenges in the literature and presents the DRAMMS algorithm to deal with them.

Chapter 4 validates the proposed DRAMMS algorithm in various registration tasks containing brain, breast and cardiac images, in cross-subject and longitudinal registration settings.

Chapter 5 applies the validated DRAMMS algorithm to the five clinical and research topics that motivate this dissertation.

Chapter 6 presents two attempts towards further improving DRAMMS speed and accuracy.

Chapter 7 concludes this dissertation and discusses the future work.

Additionally, the Appendix introduces the now publicly-available and extensively tested DRAMMS software, including the manual and the tutorials.
CHAPTER 2

Literature of Image Registration

This chapter briefly reviews the literature of medical image registration methods. More comprehensive reviews can be found in survey papers (Maintz and Viergever 1998; Lester and Arridge 1999; Hill et al. 2001; Zitova and Flusser 2003; Pluim et al. 2003; Crum et al. 2004; Holden 2008; Sotiras et al. 2012). Despite being brief, the literature review in this chapter serves three purposes: 1) to establish the state of the art in image registration; 2) to identify two common and important challenges in the literature that technically motivate the development of a new deformable registration algorithm later in this dissertation; 3) to present some registration algorithms whose implementations are publicly-available, against which the proposed method is compared in the validation part of this dissertation.

The remaining of this chapter is organized as follows: Section 2.1 formally defines image registration in mathematical terms, and formulates it as an optimization problem. Section 2.2 classifies existing approaches into two major categories (landmark/feature-based methods and voxel-wise methods), based on whether part, or all of image voxels are used to optimize the registration cost function. Sections 2.3 and 2.4 review approaches in these two categories, respectively. Finally, Section 2.5 reviews several registration approaches whose implementations are publicly available, against which our proposed method will be compared in Chapter 4 of this dissertation.
2.1. Image Registration: General Formulation

As being previously defined and visualized in Section 1.1, image registration is the process of transforming different images into the same coordinate system, so that after registration, the same spatial location in different images correspond to the same anatomical structure. Here, a more technical definition is given. It is sketched in Fig. 7. In mathematical terms, two 2-dimensional (2D) or 3-dimensional (3D) gray-scale images can be denoted as $I_1 : \Omega_1 \subset \mathbb{R}^d \mapsto \mathbb{R}$ and $I_2 : \Omega_2 \subset \mathbb{R}^d \mapsto \mathbb{R}, (d = 2, 3)$, where $\Omega_1$ and $\Omega_2$ represent two different 2D (when $d = 2$) or 3D (when $d = 3$) sub-spaces where the two images reside.

Image registration is the process of finding an optimal deformation $h$, such that voxel $x$ in image $I_1$ is spatially aligned with its corresponding voxel $h(x) : \Omega_1 \mapsto \Omega_2$ in image $I_2$. In deformable image registration, which is the focus of this dissertation, every voxel $x$ can move freely and independently if there is no regularization of the deformation. Therefore, the deformation $h$ is generally modeled as $h(x) = x + u(x)$, where $u(x)$ is the movement, or displacement, needed to deform voxel $x \in I_1$ to its anatomically corresponding voxel $h(x) \in I_2$. Such a process of deforming each and every voxel in image $I_1$ is often known as mapping, or warping, image $I_1$ by deformation $h$, or, applying deformation $h$ onto image $I_1$, which is often denoted as $h \circ I_1$. 

Figure 7: Sketch of the image registration process.
From the above definition, it should not be difficult to see that the essential issue in image registration is to accurately find correspondences across images. Usually, this is done by maximizing the similarity $S(h \circ I_1, I_2)$ between the transformed source image $(h \circ I_1)$ and the target image $(I_2)$. In other words, we assume that the two images are in anatomical correspondence when the similarity between them is maximized. On the other hand, there is also a constraint on deformation $h$, denoted as $R(h)$. The constraint implies that the deformation from one image to the other must be spatially regularized (smooth), since human anatomical structures move smoothly. By balancing the two terms by some scalar weight $\lambda$, a registration method can be generally formulated as an optimization problem below,

$$
\begin{align*}
  h^* &= \arg \max_h S(h \circ I_1, I_2) + \lambda R(h).
\end{align*}
$$

This formulation is also referred to as the "registration energy", or the "registration cost function". Almost all registration methods can be represented by this formulation.

### 2.2. Classification of Existing Registration Methods

To review the literature more systematically, this dissertation classifies existing image registration approaches into two major categories, based on whether all, or only a subset of, image voxels are used to evaluate the similarity between two images. This classification criterion is also used in the survey paper [Maintz and Viergever, 1998]. One category, \textit{landmark/feature-based methods}, first detects a subset of geometrically salient points as landmarks (\textit{e.g.}, surfaces, corners, boundary points, line intersections), establishes correspondences on them, and then propagates the correspondences throughout the entire image space. The other category, \textit{voxel-wise methods}, simultaneously find correspondences at each and every image voxel, by maximizing an overall similarity between two images. Approaches in each of the these two categories are reviewed below.
2.3. Landmark-based Registration Methods

Methods in this category are intuitive and straightforward, mimicking the way human experts register two images — extract landmarks, establish correspondences on landmarks, and propagate the correspondences to the rest of the image. Therefore, a landmark-based method usually consists of three components: landmark extraction, landmark matching, and correspondence propagation. Below are reviews of these three components, followed by an analysis of the limitations in this category of methods.

2.3.1. Landmark Extraction

Landmarks, or feature points, are voxels that either geometrically salient (such as the edges, corners and boundaries), or anatomically special (such as mass centers), or both. Manually defining landmark points in images is straightforward, but also subjective and laborious. Hence, some studies have used fiducial markers to label landmark points (e.g., Maurer et al., 1997; Kremser et al., 1997; George et al., 2011)). In these approaches, small objects that can be detected by imaging devices are usually invasively placed at some pre-defined anatomical locations. They leave visible points known as fiducial markers in images. The promise is that a registration algorithm should align anatomical structures defined by fiducial markers, as they provide absolute ground-truth for image registration. However, the drawback is its invasiveness. As computerized technologies evolve, non-invasive approaches have received more attention for landmark extraction. Various image filters (e.g., Canny (Canny, 1986), Anisotropic Diffusion (Perona and Malik, 1990), etc) have been used to detect edge or corner points as landmarks (Hsieh et al., 1997; Zana and Klein, 1999; Can et al., 2002; Livyatan et al., 2003). Texture descriptors (e.g., wavelet (Burrus et al., 1998), Gabor (Manjunath and Ma, 1996), SIFT (Lowe, 1999), image moments (Mukundan and Ramakrishnan, 1998), scale-space analysis (Lindeberg, 1998), etc.) are also used to extract geometrically distinctive
landmark points, such as mass centers and line intersections (Pajares and de la Cruz 2004; Zhan et al. 2007). Model-based or contour-based object recognition methods have also been used to extract structure/organ boundaries as landmarks for image registration (Li et al. 1995; Chen et al. 2010; Yin et al. 2012).

### 2.3.2. Landmark Matching

After having extracted landmarks from two images, the next step is to establish landmark correspondences across images. If the set of landmarks in one image is known to have one-to-one correspondences with the set of landmarks in the other image, we can establish their correspondences using the Iterative Closest Point (ICP) algorithm (Besl and McKay 1992), or its variants (Zhang 1994; Feldmar et al. 1997). The ICP algorithm establishes correspondences based on geometric distances between voxels. That is, voxels spatially close by are more likely to correspond to each other. Based on the initial estimation of landmark correspondences, a deformation is calculated. After the images have been deformed, a new closest neighbor is assigned to a landmark as its corresponding point. The deformation and the landmark correspondences are then updated iteratively until they both reach some convergence criteria. The ICP algorithm is challenged when outlier landmarks are present (an outlier landmark is the one that could not find its corresponding landmark in the other image). The presence of outlier landmarks is mainly because landmarks are extracted independently from two images without considering their corresponding relationship. To address this challenge, fuzzy correspondence algorithms have been developed (Chui and Rangarajan 2003; Yang et al. 2011). They relax the requirement for the one-to-one deterministic correspondence to one-to-many or many-to-one fuzzy correspondences, and allow outlier landmarks not to establish correspondence with any landmark in the other image. Another limitation in the ICP algorithm is its similarity criterion — the similarity between two voxels is only measured by the spatial distance between them. Actually, spatial distances do not necessarily indicate correspondences, as corresponding voxels may be
spatially far away from each other. To deal with this challenge, various similarity measures based on a landmark’s context information have been used, such as (normalized) mutual information (Zhan et al., 2007), curvature similarity (Fischer and Modersitzki, 2004) and local histogram (Shen, 1997).

2.3.3. Correspondence Propagation

This step is to propagate the correspondences from the landmark locations to the rest of the image. In the literature, this is usually done through interpolation. The interpolation is based on the assumption that the movement $u(x)$ of a non-landmark point $x$ to its correspondence is the smooth interpolation based on the distances between this non-landmark point $x$ to all $N$ landmark points $p_i (i = 1, 2, \ldots, N)$. One family of interpolation methods is the one that interpolates by the radial basis functions (RBFs) $\phi(\cdot)$ of the distances between a non-landmark point and all landmarks. Mathematically, the interpolated movement $u(x)$ at a non-landmark point $x$ is expressed as

$$u(x) = \sum_{i=1}^{N} \omega_i \phi(|x - p_i|) \quad (2.2)$$

where $\omega_i$ is the coefficients denoting the influence of each landmark point $p_i$. $\omega_i$’s can be solved in closed-form by plugging in the known movements of $N$ landmarks into the above equation. The advantage of the RBF-based interpolation is its closed-form solution, even when the landmarks are irregularly placed in the image space. One typical and widely cited/used RBF is the thin-plate-spline (TPS) function (Bookstein, 1989), and hence the so-called TPS-based deformation interpolation. It is defined as $\phi(r) = r^2 \log(r)$ in 2D and $\phi(r) = r$ in 3D. Nevertheless, there are several drawbacks in the above interpolation framework. One limitation is the global support of landmarks. Ideally, the movement of a query voxel should be influenced only by landmarks close to it. However, in the formulation
above, all landmarks (including the landmarks located far from the query voxel) influence the movement of the query voxel. To solve this problem, TPS interpolations with local supports have been proposed \cite{Fornefett2001}. Thus, a landmark only has influence in the neighboring regions. Another limitation of the TPS interpolation comes from its dependency on the exact location of landmarks. As landmarks may often have localization errors, the dependency on the exact landmark locations will inevitably propagate the landmark localization errors to the rest of the image. To solve this problem, the approximating TPS (aTPS) \cite{Rohr2001} was proposed, which allows, in the TPS interpolation framework, some anisotropic localization errors for landmarks. However, the estimation of localization errors for each landmark becomes a new issue when the ground-truth landmark locations are unknown.

2.3.4. General Remarks on Landmark-based Methods

Landmark-based registration methods are intuitive and straightforward, mimicking the way that human experts register two images. However, it is hard to find a common set of landmark extraction and landmark matching methods that can be generally applied to various registration tasks (longitudinal and cross-subjects) and to medical images of various organ (brain, breast, heart, etc). This is because the detection and matching of landmarks are often dependent on the specific geometry or anatomy in the input images. For instance, the method to detect the boundary between white matter tissue and gray matter tissue for brain registration should be different from the method to detect the boundary of ventricles in cardiac registration, because the latter boundary has a regular shape while the former boundary does not. Besides, both landmark detection and landmark matching are subject to errors that are often inevitable, which will propagate to other parts of the image during the correspondence propagation step. In addition, landmarks are often irregularly distributed in the image space. As a result, the regions having more landmarks usually have higher registration accuracy than the regions having fewer landmarks. Because of these three
drawbacks, landmark-based methods are not suitable for the five studies in Section 1.1.

2.4. Voxel-wise Registration Methods

Registration methods in this category use all image voxels to find the optimal deformation. Therefore, it is often more generalizable to various registration tasks, for there is no need to extract task-specific landmarks. A voxel-wise method usually has three key components. The first component is the **similarity metric**. The alignment of two images is based on maximizing the similarity between them. Therefore, the definition of similarity metrics largely determines the performance of image registration. The second component is the **deformation model**. A deformation model defines the way one image is deformed into the space of the other image. A deformation model can be local or global, and can have varying degrees of freedom, which influence the accuracy of a voxel-wise registration method. The third component is the optimization strategy. Numerical optimizer is needed since most image registration problems do not have a closed form solution. Approaches for these three components are reviewed separately below.

2.4.1. Similarity Metrics

In the registration of mono-modality images (**i.e.,** two images are from the same imaging modality), "Sum of Squared Differences" (SSD) of the intensities between two images is a simple similarity metric (**e.g.** [Kybic and Unser, 2003]). Given two intensity images $I$ and $J$, which have the same image lattice $\Omega$ (if not, we can re-sample one image into the same image lattice of the other), the SSD metric is defined as:

$$SSD(I, J) = \sum_{x \in \Omega} \|I(x) - J(x)\|^2. \quad (2.3)$$
Using SSD for registration is based on the assumption that the same anatomical structure exhibits the same or very similar intensities in the two images to be registered. In practice, image inhomogeneity, noise and different imaging protocols can pose challenges to this assumption.

To address these challenges, other studies have used the (normalized) correlation coefficient (CC). Without loss of generality, we can assume that two images \( I \) and \( J \) are in the same image lattice \( \Omega \) (if not, we can resample them), then the CC metric is defined as

\[
\text{CC}(I, J) = \frac{\sum_{x \in \Omega} (I(x) - \mu_I)(J(x) - \mu_J)}{\sqrt{\sum_{x \in \Omega} (I(x) - \mu_I)^2} \sqrt{\sum_{x \in \Omega} (J(x) - \mu_J)^2}},
\]

where \( \mu_I \) and \( \mu_J \) are the mean intensities in these two images. Actually, CC can be defined either globally between two entire images, which is the above equation, or locally between patches/neighborhoods of two voxels (e.g., (Kim and Fessler, 2004)). In the latter case, the only things that change in the above equation are: i) the entire image lattice \( \Omega \), which should be replaced by the neighborhood of voxel \( x \), and ii) the mean intensities of the entire images, which should be replaced by the mean intensities of the neighborhoods of voxel \( x \) in the two images. The use of CC for registration is based on the assumption that two images, or patches, are spatially aligned when their intensity distributions are highly correlated. The CC metric can be used in both mono- and multi-modality registration tasks. The study in (Klein et al., 2009) shows that the CC metric has higher accuracy and increased level of robustness to intensity inhomogeneities compared with the SSD similarity metric.

Another widely-used similarity metric is mutual information (MI) (Wells et al., 1996; Maes et al., 1997). It is borrowed from the information theory. Treating two images as two signal sequences, mutual information describes how much intensity information we know about one image if the intensity information about the other image is completely given. Given
two images $I$ and $J$, the MI metric is defined as

$$
\text{MI}(I, J) = H(I) - H(I|J)
$$

$$
= H(J) - H(J|I).
$$

Here $H(\cdot)$ is the entropy of an image, or, the degree of chaos in the intensity distribution of an image. $H(\cdot|\cdot)$ is the conditional entropy, or the degree of chaos in the intensity distribution of image $\cdot$ when the intensity distribution of image $\cdot$ is completely given. When two images are in a complete anatomical alignment, the information of one image can be inferred with least amount of uncertainty (i.e., minimum chaos), given the information of the other image. In this case, the MI metric reaches its maximum. Conversely, maximizing mutual information is assumed to maximize the anatomical alignment of the two input images. The MI metric can also be used in both mono- and multi-modality registration tasks, showing robustness with regard to intensity inhomogeneities, image contrast differences, and the partial volume effects \cite{Pluim2003}.

All metrics mentioned above are intensity-based similarity metrics. Some approaches have characterized voxels with richer information than intensities, often by taking into account the information in the neighborhood of voxels. Richer information characterizes voxels more distinctively and results in higher accuracy in establishing correspondences. One example is the geometric information described by image moments \cite{Shen2002}. Other examples include texture information extracted by wavelet or Gabor texture descriptors \cite{Xue2004, Liu2002}. While this direction shows great promise, there is also a need for develop a distinctive and general geometric/texture descriptor for voxels. Distinctiveness in voxel characterizations is the key for accurate matching; while generality is preferred so that the descriptor can be applied to as many registration tasks as possible. One major contribution of the method proposed in this dissertation is an attempt to satisfy the need mentioned above.
2.4.2. Deformation Models

Deformation models are either physical or geometric. They are defined and reviewed separately below.

Physical Deformation Models

Physical models are usually governed by partial differential equations (PDEs) that can find their roots in physical phenomena. Examples include elastic models (Broit, 1981; Bajcsy and Kovacic, 1989; Davatzikos, 1997; Gee and Bajcsy, 1999), viscous fluid flow models (Christensen et al., 1996; DAgostino et al., 2003), diffusion models (Thirion, 1998) and diffeomorphic models (Joshi and Miller, 2000; Beg et al., 2005; Vercauteren et al., 2009).

Elastic models originate from the Navier-Cauchy equation. They model an image grid as an elastic membrane. The deformation is controlled by two competing forces until an equilibrium is reached — the external force deforms the image while the internal force preserves elasticity of the structure/material (Broit, 1981; Bajcsy and Kovacic, 1989; Davatzikos, 1997; Gee and Bajcsy, 1999). One drawback in elastic models is that they cannot handle large deformations. To handle large deformations, viscous fluid flow models are developed (e.g., Christensen et al., 1996; DAgostino et al., 2003). They operate on the so-called velocities to deform images. The velocities are updated in small time intervals, and can accumulate over time to cover large deformations (Christensen et al., 1996; DAgostino et al., 2003). On the other hand, small-step increments cause computational inefficiency, which is the main drawback of viscous fluid flow models. To speed up the process, a relatively simpler model, diffusion model, is often used (e.g., (Thirion, 1998)). Diffusion models connect the diffusion equation to a Gaussian function, which is used to smooth the deformation. The Demons deformation (Thirion, 1998) is an example (more details about the Demons algorithm can be found later in 2.5). However, diffusion models do not necessarily guarantee smoothness, invertibility, or differentiability, which are desirable properties of deformations. To obtain these nice properties, diffeomorphic models are proposed. One example is the diffeomorphic...
version of the Demons algorithm \cite{Vercauteren2009}, which is explained in more detail in Section 2.5. Another example of using diffeomorphic models is in the Large Deformation Diffeomorphic Metric Mapping (LDDMM) algorithm \cite{Joshi2000,Beg2005}. Because of its theoretical contribution and its popularity, LDDMM is reviewed in more detail here.

LDDMM assumes that the deformation $\phi$ that deforms the source image $I_1$ into the space of the target image $I_2$ is diffeomorphic (smooth, one-to-one, invertible, differentiable, and inversely differentiable). The diffeomorphic deformation $\phi$ can be parameterized by spatial coordinate $x$ and time $t$, \textit{i.e.}, $\phi = \phi(x, t)$. It can be further generalized by integrating a time-dependent smooth velocity field $v$ through an ordinary differential equation (ODE),

$$
\frac{d\phi(x, t)}{dt} = v(\phi(x, t), t) := v(x, t) \quad t \in [0, 1]. \tag{2.6}
$$

The deformation $\phi$ is the accumulation of the velocity field $v(\cdot, t)$ from time $t = 0$ to time $t = 1$. At time $t = 0$, images are not yet deformed, and the deformation is an identity, \textit{i.e.}, $\phi(x, 0) = x$. At time $t = 1$, the images are deformed by $\varphi$ into anatomical correspondences, \textit{i.e.}, $\phi(x, 1) = \varphi$. The key problem is to compute the velocity $v$. This is done in the following optimization formulation:

$$
\arg\min_v \left( \|\varphi^{-1} \circ I_1 - I_2\|_{L_2}^2 + \lambda \int_0^1 \|v(x, t)\|_{V}^2 dt \right). \tag{2.7}
$$

In this formulation, the first term is the similarity between the deformed and the target images. The similarity is measured by the $L_2$ norm of the intensity difference between two images. The second term, weighted by scalar $\lambda$, is the regularization term, which is the main feature of LDDMM. The velocity field is assumed to reside in the Hilbert space $V$, \textit{i.e.}, $v(x, t) \in V$. Therefore, we can define the norm of a velocity field as $\|v(x, t)\|_V$, which is further defined as

$$
\|v(x, t)\|_V = \|Lv(x, t)\|_{L_2}. \tag{2.8}
$$
In this definition, $\|\cdot\|_{L_2}$ is the standard $L_2$ norm, and $L$ is a differential operator to guarantee existence of solutions in the space of diffeomorphism (the definition of operator $L$ is given in Beg et al. (2005)). The definition of the norm of the velocity field in Eq. 2.8 makes the LDDMM cost function in Eq. 2.7 a mathematically rigorous formulation, in the sense that the solution of Eq. 2.7 guarantees the desirable diffeomorphism in the obtained deformation (the mathematical proof can be found in Beg et al. (2005)).

Despite its mathematical elegance, LDDMM suffers from two limitations. One is the computational inefficiency, because calculating the norm of velocity fields in the Hilbert space is computationally costly. Another limitation is that it is not symmetric with regard to the input images, as pointed out in Avants et al. (2008). To solve this problem, a symmetric diffeomorphic algorithm is developed in Avants et al. (2008), and is included in the ANTs registration framework that will be separately reviewed later in Section 2.5.

**Geometric Deformation Models**

Besides physical models, geometric models are also used in image registration. In geometric models, the movement of one voxel is often smoothly interpolated by the movements of neighboring voxels. Geometric models are often more flexible, computationally more efficient, because there is no need for numerically solving partial differential equations as in physical models. One representative geometric model is the Free-Form Deformation (FFD) model Rueckert et al. (1999), which has been one of the most commonly used deformation model in the medical image registration community (as evidenced by more than 2,300 citations since its publication in 1999). In the following, the FFD model is reviewed in more detail as it is the model that will be used in this dissertation.

Fig. 8 shows the idea of FFD model. For simplicity, the figure only shows a 2D FFD; a 3D FFD follows the same principle, just with one more dimension. In the FFD model, a regular grid of so-called "control points" is superimposed on top of the dense image lattice (by "dense image lattice" we mean the grid of image voxels, such as the green lattice in Fig. 8). Also in Fig. 8 control points are labeled as blue. The control points can be regarded
as regular samplings of the dense image lattice. A FFD model basically states that the movement $u(x, y)$ of an image voxel $(x, y)$ (e.g., the point under red star in Fig. 8) is a smooth, B-spline-based interpolation of the displacement $d_{i,j}$ of neighboring control points (4 neighboring control points in each dimension, i.e., $d_{i+m,j+n}$ for $m = 0 \ldots 3, n = 0 \ldots 3$). Mathematically, the equation below connects the movement of an image voxel $(x, y)$ with the displacements of its $4 \times 4$ neighboring control points:

$$u(x, y) = \sum_{m=0}^{3} \sum_{n=0}^{3} B_m(p)B_n(q)d_{i+m,j+n} \quad (2.9)$$

where $(i = \lfloor x/n_x \rfloor - 1, j = \lfloor y/n_y \rfloor - 1)$ is the starting control point that affects the current ordinary point $(x, y)$, with $n_x$ and $n_y$ being the numbers of voxels between adjacent control points in x and y dimensions. $p = x/n_x - \lfloor x/n_x \rfloor$ and $q = y/n_y - \lfloor y/n_y \rfloor$ are interpolation hyper-parameters determined by the relative distance between the ordinary point and the
control point that originates the closest control point box around this ordinary point. Those hyper-parameters are used in B-spline basis functions $B_0(\cdot), B_1(\cdot), B_2(\cdot), B_3(\cdot)$ to determine the final interpolation weights. In each dimension, four B-spline basis functions are used to represent the influence of four neighboring control points in this dimension to the current ordinary point. The influence is based on spatial distance — spatially closer control points have higher levels of influence to the current ordinary point. Mathematically, the B-spline basis functions are

\begin{align*}
B_0(a) &= (1 - a)^3/6 \\
B_1(a) &= (3a^3 - 6a^2 + 4)/6 \\
B_2(a) &= (-3a^3 + 3a^2 + 3a + 1)/6 \\
B_3(a) &= a^3/6 \\
\end{align*}

The advantages of the FFD model are two-folds. First, computational efficiency: as a result of this mathematically rigorous model (Eq. 2.9), the task of finding movements for each and every image voxel has been translated to finding displacements for at a much smaller number of regularly sampled control point locations. Second, the intrinsic smoothness of the obtained deformation, which is naturally brought by the smooth and differentiable B-spline basis functions (Eq. 2.10). Because of these advantages, FFD model is used in the proposed registration algorithm in Chapter 3 of this dissertation.

2.4.3. Optimization Strategies

Because most registration problems do not have closed-form solutions, the task of finding the optimal values for deformation parameters is left to a numerical optimizer. Numerical optimizers, or optimization strategies, can be either continuous or discrete. They are defined
Continuous optimizers assume that the registration cost function is continuously valued and is differentiable with regard to the parameters in deformation models. Therefore, continuous optimizers can travel on the continuous curve or surface of the registration cost function, and find the optimal parameter values that lead to the extrema of the registration cost function. In general, continuous optimizers are intuitive, easy to implement, guaranteed to converge (at least to local minima), and have shown versatility and accuracy in numerous image registration approaches. Some typical continuous optimizers are: Gradient Descent strategy (Cauchy 1847), Newton strategy (Avriel 2003), Powell’s strategy (Powell 1964), Levenberg-Marquardt strategy (More 1978), and Stochastic Gradient strategy (Tsitsiklis et al. 1986). Interested readers can refer to (Klein 2008) for an in-depth study of different optimization strategies in the context of medical image registration. Despite the popularity and success, continuous optimizers are known to have several drawbacks, including being sensitive to the initial condition and local minima, and being computationally inefficient.

Compared to continuous optimizers, the research of using discrete optimization strategies in the context of image registration is more recent (Glocker et al. 2008, 2011). Discrete optimizers first discretize, or sample, the entire search range of deformation parameters, and then find the sampled parameter values that lead to the extrema of the registration cost function. Compared with continuous optimizers that often perform local searches, the search in discrete optimization is global. Because of its global manner, it is believed to be less sensitive to local minima, although this belief is not yet supported by a rigorous proof. Also, discretization of the search space reduces the computation cost as the registration cost function is only computed at a limited number of sampled parameter values other than continuously at many more values in the continuous optimization. On the other hand, one may argue that the optimal values for deformation parameters may be missed from the sampling of the search space in the first place. This is partly alleviated by multi-resolution search strategies, but still remains an open issue for further investigations.
In this dissertation, we have chosen the discrete optimizer to find the optimal values for deformation parameters in the proposed registration algorithm. This is mainly because of its computational efficiency and its demonstrated accuracy in (Glocker et al., 2008). More detail of the formulation and the implementation of the discrete optimizer can be found in a later chapter (Section 3.3.2).

2.4.4. General Remarks on Voxel-wise Methods

Voxel-wise methods have gained popularity because of the simplicity in formulation, the generality to different registration tasks, and the independence from user initializations or interventions. Because of these properties, most general-purpose image registration methods (methods that can generally apply to different registration tasks) are voxel-wise methods. However, there are still several challenges for voxel-wise methods. One challenge stems from the fact that most voxel-wise methods find correspondences based on image intensities in the image. As intensity information alone does not necessarily indicate anatomical or geometrical context at a voxel, intensity-based matching usually suffers from ambiguities in finding anatomical correspondences. A second challenge arises in the presence of pathologies in patients’ images, whose correspondences are not present in the images of healthy subjects. Since voxel-wise methods use all voxels equally in the registration process, the pathological regions/voxels will bring about negative impact to image registration. Towards the solution of these two challenges, a new image registration method is developed in the next chapter.

2.5. Some Publicly-Available Image Registration Tools

Among the vast number of image registration methods developed in the literature, some are fully implemented and publicly released to the image registration community. They
become valuable tools for promoting future technical innovations and for facilitating large-scale research and clinical studies. This section reviews some of those publicly-available image registration tools. They are all voxel-wise methods that can be generally applied to a variety of image registration tasks. They all contain complete pipelines of similarity metrics, deformation models and numerical optimizers, which are the three key components in a voxel-wise registration method. They make contributions in one or more of these components. They are also closely related to the method that is developed in this dissertation. As we will see in Chapter 4, some of them are chosen as references to validate the developed algorithm in this dissertation.

The image registration tools reviewed in this section include: AIR (Woods et al., 1992, 1998), ART (Ardekani et al., 2005), ANTs (Avants et al., 2008), (Diffeomorphic) Demons (Thirion, 1998; Vercauteren et al., 2009), DROP (Glocker et al., 2008), flirt (Jenkinson and Smith, 2001), fnirt (Andersson et al., 2007, 2008), and IRTK (Rueckert et al., 1999). In addition, two examples of public tools developed in the past 2-3 years are mentioned in the end of this section (elastix (Klein et al., 2010) and NiftiReg (Modat et al., 2010)). It should be noted that this is only an incomplete list of publicly-available voxel-wise registration tools in the literature. However, they represent a wide variety of innovations and choices in the similarity metrics, deformation models and numerical optimizations.

AIR

The Automatic Image Registration (AIR) registration tool was developed by researchers (R. Woods et al.) at the University of California, Los Angeles (UCLA) in the 1990s (Woods et al., 1992, 1998). AIR makes contributions in the deformation models. It models the deformation by second-, third-, fourth- and fifth-order non-linear polynomials (30, 60, 105, and 168 deformation parameters, respectively). Eq. 2.11 below is the second-order polynomial that is used in AIR to parameterize a deformation. Given a voxel located at coordinate \(x, y, z\) in one image, the coordinate \((x', y', z')\) of its corresponding voxel in the other image is modeled by

\[
\quad
\]
Here $P$’s are the deformation parameters to be optimized during AIR registration. Higher-order polynomials are analogous to the second order model above but involve more deformation parameters.

To find the optimal values for the deformation parameters, AIR minimizes a cost function between the deformed image and the target image. Three different cost functions are supported in AIR. One is the ratio image uniformity (RIU). The ratio of the intensity in the deformed image to the intensity in the target image is computed at each voxel location. The deformed and the target images are assumed to be aligned when the ratio is homogeneous (i.e., high mean value and low standard deviation) in the entire target image space. The second cost function is Sum of Square Difference (SSD) of image intensities. The SSD cost function assumes that the same anatomical structure in different images share the same intensity. Therefore, two images are aligned when the differences in their intensities are minimized. The third cost function is a variant of the second cost function, with some relaxation. Instead of assuming that the same anatomical structure shares the same intensity in different images, the third cost function assumes the same anatomical structure shares the same intensity with a global scaling factor. Therefore, it is a globally-weighted SSD cost function. To minimize the cost function, a numerical optimizer based on either Newton’s method [Avriel 2003] or Levenberg-Marquardt method [More 1978] is used.

The AIR registration tool is publicly available at [http://bishopw.loni.ucla.edu/air5/](http://bishopw.loni.ucla.edu/air5/).
ART

The Automatic Registration Toolbox (ART) was developed by researchers (B. Ardekani et al) at the Nathan Kline Institute for Psychiatric Research and in the New York University in 2005 (Ardekani et al., 2005). It makes contributions in defining a new similarity metric. In the ART framework, each voxel $x$ is characterized by a high-dimensional feature vector $f(x)$. The feature vector $f(x)$ is constructed by stacking the intensity values of all the voxels in the neighborhood of the current voxel $x$. Compared with most voxel-wise methods which establish correspondences by the intensity at each voxel, ART establishes correspondences by the high-dimensional feature vector at each voxel. The similarity of two voxels $x$ and $y$ is defined on their feature vectors as follows

\[
\text{sim}(x, y) = \frac{f(x)^T H f(y)}{\sqrt{f(y)^T H f(y)}} = \frac{(H f(x))^T (H f(y))}{\sqrt{(H f(y))^T (H f(y))}}
\]

(2.12)

where $H$ is an idempotent (i.e., $H^2 = H$) and symmetric ($H^T = H$) matrix that removes the mean of the vector it pre-multiplies.

ART is implemented in multi-resolution fashion. In the middle and low image resolutions, ART searches correspondences for all voxels in the target image. In the highest image resolution, ART only searches correspondences at those voxels whose gradient norms are in a certain upper percentile of the gradient magnitude histogram. The obtained deformation is smoothed by a Gaussian filter to regularize that the voxels close to each other remain close after the image is deformed.


ANTs

The Advanced Normalization Tools (ANTs) was developed by researchers (B. Avants et al)
at the University of Pennsylvania in the 2000s \cite{Avants2008}. It is based on the LD-DMM algorithm (reviewed earlier in Section 2.4), but improves LDDMM’s computational efficiency and introduces symmetry into the LDDMM framework.

**ANTS** decomposes the diffeomorphic deformation $\phi$ into two components $\phi_1$ and $\phi_2$. The idea is that, instead of deforming image $I_1$ into the space of image $I_2$ with one deformation $\phi$, which is not symmetric to the input images, ANTs simultaneously deforms image $I_1$ with deformation $\phi_1$ and image $I_2$ with deformation $\phi_2$, each towards the ”midpoint” image between $I_1$ and $I_2$. Therefore, the formulation becomes symmetric to the input images. We can define, in $t \in [0, 0.5]$, $v(x, t) = v_1(x, t)$, and in $t \in [0.5, 1]$, $v(x, t) = v_2(x, 1 - t)$, and at the midpoint (when $t = 0.5$), the deformation in both components are $\phi_1 = \phi_1(x, 0.5)$ and $\phi_2 = \phi_2(x, 0.5)$. This leads to a symmetric variant of the LDDMM’s energy function in Eq. \ref{eq:LDDMM} as follows:

$$\arg \min_{v_1, v_2} \left( \| \varphi_1^{-1} \circ I_1 - \varphi_2^{-1} \circ I_2 \|_2^2 + \lambda \left( \int_0^{0.5} \| v_1(x, t) \|_V^2 dt + \int_0^{0.5} \| v_2(x, t) \|_V^2 dt \right) \right). \tag{2.13}$$

By this formulation, ANTs achieves symmetric diffeomorphic deformations. Also, ANTs supports three intensity-based similarity metrics: SSD, which LDDMM uses, and which is shown in the above Eq. \ref{eq:LDDMM} for simplicity of the ANTs formulation; CC, which is used by default in ANTs; and MI. The gradient descent optimization strategy \cite{Cauchy1847} is used to numerically find the optimal deformation in the above formulation.

The ANTs registration tool is publicly available at \url{http://www.picsl.upenn.edu/ANTS}.

**Diffeomorphic Demons**

The Demons algorithm was first proposed by J. P. Thirion at INRIA, Equipe Epidaure, France, in 1998 \cite{Thirion1998}. It was later formulated into the diffeomorphic version and
Thirion’s Demons algorithm considers deformable image registration as a non-parametric diffusion process. He introduces Demons, to push voxels to their correspondences according to the local intensity characterizations. A force is computed from the optical flow equations to push voxels by some displacement that is iteratively added to the total displacement (initially zero). The total displacement is then smoothed with a Gaussian filter serving as regularizations of the deformation. Below is the equation to compute velocity $v$. Given two images $I_1$ (the moving image) and $I_2$ (the fixed image), and the current deformation $h$, Demons updates the deformation by velocity $v$, which is computed as follows:

$$
v = \frac{(I_2 - h \circ I_1) \nabla I_2}{\| \nabla I_2 \|^2 + \alpha ((I_2 - h \circ I_1) \nabla I_2)^2}. \quad (2.14)
$$

The numerator part of the above equations shows that the similarity metric in Demons is the sum of intensity difference. Ideally one would expect that when the image gradient $\nabla I_2$ is close to 0, the velocity should also be close to 0. Therefore, the term $\alpha (I_2 - h \circ I_1)^2$ is introduced into the denominator to guarantee the stability of the computed velocity field.

The computed velocity field $v^n$ in the $n^{th}$ iteration is used to update the current deformation $h^n$ into a new deformation $h^{n+1}$, which initializes the next iteration. There are three different ways to update the deformation:

**Additive Demons:**

$$h^{n+1} = h^n + v^n \quad (2.15)$$

**Compositive Demons:**

$$h^{n+1} = h^n \circ (\text{Identity} + v^n) \quad (2.16)$$

**Diffeomorphic Demons:**

$$h^{n+1} = h^n \circ \exp(v^n) \quad (2.17)$$
Thirion’s Demons uses the additive model. However, it does not ensure diffeomorphism, nor
does the Compositive Demons. Therefore, the diffeomorphic Demons version is proposed in
(Vercauteren et al., 2009), and is shown to produce much smoother deformation (measured
by Jacobians and Harmonic Energy of the obtained deformation) at an accuracy comparable
to that of the Additive Demons.

The Demons registration tool (including the implementation of additive, compositive and
diffeomorphic versions) is publicly available at http://www.insight-journal.org/browse/
publication/154

DROP

DROP is the implementation of a deformable image registration pipeline developed by
researchers (B. Glocker et al) at the Ecole Centrale de Paris, France, the Technische Uni-
versität München, Germany, and the University of Grete, Greece (Glocker et al., 2008). The
main contribution of DROP is in the numerical optimization. Discrete optimization is, for
the first time, introduced into the field of medical image registration, bringing in significant
speedup compared with other continuous optimizers such as gradient descent (8 minutes
by discrete optimization versus 3 hours 50 minutes by gradient descent for the same set of
brain images, as reported in (Glocker et al., 2008)). The discrete optimization strategy is
briefly reviewed earlier in this chapter (Section 2.4) and is elaborated in Section 3.3.2 of
this dissertation.

DROP uses FFD as its deformation model, and supports 12 intensity-based similarity met-
rics: Sum of Absolute Differences (SAD), which is used by default, Sum of Absolute Dif-
ference plus Sum of Gradient Inner Products (SADG), Sum of Squared Differences (SSD),
Normalized Correlation Coefficient (NCC), Normalized Mutual Information (NMI), Corre-
lation Ratio (CR), Sum of Gradient Inner Product (SGAD), and others.

The DROP registration tool is publicly available at http://www.mrf-registration.net/
The FMRIB’s Linear Image Registration Tool (flirt) was developed by researchers (M. Jenkinson and S. Smith) at the University of Oxford, UK, in the early 2000s ([Jenkinson and Smith 2001](http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FLIRT)). Unlike other deformable registration methods reviewed in this section, flirt is an affine registration method. It assumes that all the voxels in a 3D image move together by a global affine transformation that has 12 degrees of freedom (dof), including 3 dof for scaling, 3 dof for translation, 3 dof for rotation, and 3 dof for shearing. Less degrees of freedom are also supported, such as in the rigid-body transformation. Because of its contributions in the optimization strategy, and because of its demonstrated accuracy and robustness in mono- and multi-modality affine registration tasks, flirt is widely cited in the image registration community, as evidenced by more than 1500 citations in the past 11 years in the Google Scholar search engine.

The authors found that local optimization strategies (such as the gradient descent strategy) together with the standard multi-resolution implementation are not sufficient to find global optima. To address this problem, they propose a global, multi-start and multi-resolution optimization strategy specifically for affine image registration problems. The fundamental idea is to combine a fast local optimization (Powell’s conjugate gradient descent method ([Powell 1964](http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FLIRT))) with an initial search phase. Initially, 8-mm cubed voxels are used and a full search is conducted over all rotation angles. Following this, various local optimizations are performed with a variety of starting points in the local neighborhood of the best points identified in the search. These local optimizations are done using 4, 2 and finally 1mm cubed voxels ([Jenkinson and Smith 2001](http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FLIRT)).

The flirt pipeline supports a wide variety of intensity-based similarity metrics, including Least Square Intensity Difference (LS), (Normalized) Correlation Coefficient ((N)CC) and (Normalized) Mutual Information ((N)MI).

The flirt registration tool is publicly available at [http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FLIRT](http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FLIRT).
fnirt

The FMRIB’s Non-linear Image Registration Tool (fnirt) was developed by researchers (J. Anderson, S. Smith and M. Jenkinson) at the University of Oxford, UK, in 2007 [Andersson et al., 2007, 2008]. The fnirt method uses Sum of Squared Differences (SSD) as the similarity metric, therefore it is only suitable for mono-modality image registration tasks. It implements the free form deformation (FFD) model. The deformation is regularized by the magnitude of the Laplacian of the deformation (also known as the bending energy of the deformation). The main contribution is in the optimization process [Andersson et al., 2007]. The optimization is based on multi-resolution Levenberg-Marquardt strategy [More, 1978]. The registration is initialized and run to the convergence in the down-sampled images, generating a deformation field with low resolutions and a high regularization weight. The images and the deformation field from the first step are then up-sampled, with the regularization modified. And it is again run to the convergence. This is repeated until the required high-resolution and the required level of regularization is achieved.

The fnirt registration tool is publicly available at [http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FNIRT](http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FNIRT).

IRTK

The Image Registration Toolkit (IRTK) was developed by researchers (D. Rueckert et al) at the Imperial College London, UK, in the late 1990s [Rueckert et al., 1999]. The main contribution is the mathematically elegant definition of the free form deformation (FFD) model, which can be combined with various similarity metrics and various optimization strategies in medical image registration field. Details of the FFD model have already been reviewed earlier in this chapter (Section 2.4).

The IRTK implementation supports three intensity-based similarity metrics: SSD, CC and MI. They are all optimized by the gradient descent optimizer [Cauchy, 1847] in multi-resolution fashion.
The IRTK registration tool is publicly available at http://www.doc.ic.ac.uk/~dr/software/.

Recent Trends

In addition to the ones mentioned above in this section, active research in the medical image registration field has led to several newly released registration tools in the past 2-3 years.

One example is elastix (Klein et al., 2010), developed by researchers (S. Klein, M. Staring, and J. Pluim) at the Image Sciences Institute, Universitair Medisch Centrum Utrecht, the Netherlands, and released in 2010. Publicly available at http://elastix.isi.uu.nl/, the elastix software package provides a wide variety of image registration algorithms for users to choose in specific registration tasks. It is also a valuable tool to study how different approaches (similarity metrics, especially optimization strategies) could influence image registration accuracy in various image registration tasks.

Another example is NiftiReg (Modat et al., 2010), developed by researchers (M. Modat et al.) in the University of College London, UK, and released in 2010. The NiftiReg software package implements the FFD-based image registration method (which is reviewed in Section 2.4) in a parallel framework suitable for GPU-based execution. Publicly available at http://www0.cs.ucl.ac.uk/staff/m.modat/Marcs_Page/Software.html, NiftiReg is an example of migrating image registration algorithms into the GPU-based parallel computing for the improvement in speed, which is important to meet the growing needs in large-scale clinical and research studies.
CHAPTER 3

DRAMMS Algorithm

This chapter proposes a new deformable image registration algorithm named DRAMMS — Deformable Registration via Attribute-Matching and Mutual-Saliency Weighting. In the following, Section 3.1 identifies two common challenges in the literature that render most registration methods inappropriate for the five studies introduced in Chapter 1.1. To cope with these challenges, the DRAMMS algorithm is proposed in Section 3.2. In essence, DRAMMS is a new similarity measure to establish reliable correspondences and to deal with missing correspondences. To construct a complete registration pipeline, the newly proposed similarity measure has to be combined with a proper deformation model and an effective optimization strategy. This is described in the DRAMMS implementation in Section 3.3. Then, Section 3.4 presents qualitative and quantitative experimental results in registration tasks involving brain, cardiac, breast and prostate images. The results in Section 3.4 are meant as a proof-of-concept for the generality, accuracy and robustness of the proposed DRAMMS algorithm. More extensive validations will be provided later in Chapter 4. Finally, Section 3.5 discusses many aspects of the DRAMMS framework and concludes this chapter. The content in this chapter has been published in a conference proceeding paper (Ou and Davatzikos, 2009) and later an invited journal paper (Ou et al., 2011).
3.1. Technical Motivations

As mentioned in the literature review chapter (Chapter 2), voxel-wise registration methods have higher levels of generality because they do not require task-specific landmarks or feature points. However, as pointed out below, most voxel-wise methods have two limitations. Overcoming the two limitations are the technical motivations of the proposed DRAMMS work.

The first major limitation in most existing voxel-wise registration methods is that they characterize voxels by attributes that are not necessarily optimal. Since the matching between a pair of voxels is based upon the attributes characterizing them, suboptimal attributes will unavoidably lead to ambiguities in matching (Shen and Davatzikos, 2002; Xue et al., 2004; Wu et al., 2006; Liao and Chung, 2009). Indeed, most voxel-wise methods characterize voxels only by image intensities, which alone do not necessarily contain anatomic information for establishing accurate correspondences. For instance, hundreds of thousands of gray matter voxels in a brain image share similar intensities; but they belong to different anatomical structures. To reduce matching ambiguities, other methods (e.g., Shen and Davatzikos, 2002) attempt to characterize voxels by a richer set of attributes, such as surface curvatures and the tissue membership. These attributes, although capable of reducing matching ambiguities, are often task- and parameter- specific. Therefore, they are not generally applicable for diverse registration tasks, such as in our motivational studies in Section 1.1. In this dissertation, we argue that, an ideal set of attributes should satisfy two conditions: 1) being discriminative. That is, the attributes of two voxels should be similar if, and only if, they are anatomically corresponding to each other, thereby minimizing the ambiguities in matching; 2) being generally applicable. That is, the way to extract a distinctive set of attributes should be the same in almost all medical images, regardless of the image modalities or content. Finding a set of attributes that satisfy these two criteria and developing a method to select the best components, thus become the first major contribution in the
The need for weighting voxels differently when there is partial loss of correspondence caused by pathologies. The lesion (white region) in the subject image could not find correspondences in the template from a healthy subject. Therefore, the pathological regions should be assigned with low weights to reduce the negative impact to registration.

The second major limitation in most voxel-wise image registration methods is that they often utilize all imaging data equally, which usually undermines the performance of the optimization process. Actually, different anatomical regions/voxels usually have different abilities to establish correspondences across images. An obvious example is in the presence of pathologies (e.g., lesion, tumors) in the patient images whose correspondence do not exist in the images of normal subjects. In Fig. 9, for instance, the lesion regions in the diseased brain image (subject) could not find correspondences in the normal brain image (template). In a more general setting, even if no pathologies are present, certain parts of the anatomy are still more easily identifiable than others. In Fig. 10 for instance, three voxels (depicted by red, blue and orange crosses) are examined. A similarity map is generated by calculating the attribute-based similarity between one of these three voxels in the subject image and all voxels in the template image. The similarity maps (Fig. 10(c,d,e)) reveal that a smaller number of candidates in the template image are similar to the voxel under the red cross in the subject image, more candidates are similar to the voxel under the blue cross in the subject image, and many more candidates for the voxel under the yellow star.
Figure 10: The need for weighting voxels differently in a typical image registration scenario. Similarity maps (c-e) are generated between one specific voxel (depicted by red, blue and orange crosses) in the subject image (a) and all voxels in the template image (b). The red point has higher ability to establish reliable correspondences, and hence should be assigned with a higher weight during registration, followed by the blue point and the orange point in descending order. This figure is better viewed in color version.

This indicates their different abilities to establish reliable correspondences, or conversely, different degrees of matching ambiguities. Naturally, these three points should be treated with different levels of "confidences" during registration. Unfortunately, most voxel-wise methods, such as the mutual-information based methods, treat all voxels equally, ignoring such differences. Other approaches attempt to address this issue, by driving registration using certain anatomically salient regions. However, in their approaches only a few voxels are finally utilized, and the potential contributions from other voxels are ignored. Moreover, the determination of which voxels to be utilized is largely heuristic or dependent on some a priori knowledge. In this dissertation, we propose and demonstrate that, an ideal optimization process should utilize all voxels, but with different weights, reflecting different levels of confidence for them to establish reliable correspondences across images. This is important for increasing the matching reliability in general, and for minimizing the negative impact of missing correspondences when they are present (such as the lesion region in Fig. 9(a)).
3.2. DRAMMS Methods

DRAMMS – Deformable Registration via Attribute Matching and Mutual-Saliency Weighting – is proposed to overcome these two aforementioned limitations. To overcome the first limitation, DRAMMS extracts a rich set of multi-scale and multi-orientation Gabor attributes at each voxel and automatically selects the optimal attributes. The optimal Gabor attributes render it relatively robust in attribute-based image matching, and are also constructed in a way that is readily applicable to various image content and image modalities. To overcome the second limitation, DRAMMS weights all voxels during the optimization process, based on a function referred to as ”mutual-saliency”. The automatically-calculated mutual-saliency value quantifies the matching reliability between a pair of voxels. As a result, voxels that can establish reliable correspondences between images are used with higher weights. On the other hand, voxels, such as those in the pathological regions, are used with low or even close-to-zero weights.

3.2.1. Overall Formulation

Given two images, $I_1 : \Omega_1 \rightarrow \mathbb{R}$ and $I_2 : \Omega_2 \rightarrow \mathbb{R}$, in 3D image domains $\Omega_i (i = 1, 2) \subset \mathbb{R}^3$, DRAMMS seeks a transformation $h$ that maps every voxel $x \in \Omega_1$ to its corresponding voxel $h(x) \in \Omega_2$, by minimizing an overall cost function $E(h)$,

$$
\min_h E(h) = \int_{x \in \Omega_1} ms(x, h(x)) \cdot \left( \frac{1}{d} \cdot \| A_1^*(x) - A_2^*(h(x)) \|^2 \right) dx + \lambda R(h). \tag{3.1}
$$

In the above cost function, $A_i^*(x)$ ($i = 1, 2$) is the optimal attribute vector that reflects the geometric and anatomical context around voxel $x$, and $d$ is the dimension of the optimized attribute vector. By minimizing $\frac{1}{d} \cdot \| A_1^*(x) - A_2^*(h(x)) \|^2$, we seek a transformation $h$
that minimizes the differences on the attributes of two voxels \( x \in \Omega_1 \) and \( h(x) \in \Omega_2 \). The extraction of attributes and selection of optimal attribute vector \( A^*_i(\cdot) (i = 1, 2) \) are detailed in Sections 3.2.2.1 and 3.2.2.2.

\( ms(x, h(x)) \) is a continuously-valued mutual-saliency weight between two voxels \( x \in \Omega_1 \) and \( h(x) \in \Omega_2 \) – higher reliability of their matching in the neighborhood indicates a higher mutual-saliency value, and hence higher weight in the optimization process. The definition of the mutual-saliency metric is explained in Section 3.2.3.

\( R(h) \) is a smoothness/regularization term usually corresponding to the response to the Laplacian operator, or the bending energy (Bookstein, 1989), of the deformation field \( h \), whereas \( \lambda \) is a balancing parameter that controls the extent of the smoothness of the obtained deformation.

Corresponding to the cost function in Eq. 3.1, the framework of DRAMMS is also shown in Fig. 11. The two major components – attribute matching (AM) and mutual-saliency (MS) weighting – are presented in the subsequent sections 3.2.2 and 3.2.3.

3.2.2. Attribute Matching (AM)

This section addresses the first major component of DRAMMS, namely, attribute matching (AM). The aim is to extract and select the optimal attributes that reflect the geometric context of each voxel. At the same time, the attribute extraction and selection pipeline should
be readily applicable to diverse registration tasks. This component consists of two modules: attribute extraction (described in section 3.2.2.1) and attribute selection (described in section 3.2.2.2).

### 3.2.2.1. Attribute Extraction

The idea of characterizing each image voxel with a rich set of attributes has been previously explored in the community. In the pioneering work (Shen and Davatzikos, 2002), the authors incorporated geometric-moment-invariant (GMI) attributes, tissue membership attributes and boundary/edge attributes into a high-dimensional attribute vector to better characterize a voxel in human brain images. Since then, a number of geometric and texture attribute descriptors have been used in image registration, including, but not limited to, intensity attributes (e.g., (Ellingsen and Prince, 2009; Foroughi and Abolmaesumi, 2005)), boundary/edge attributes (e.g., (Shen and Davatzikos, 2002; Zacharaki et al., 2008; Sundar et al., 2009; Ellingsen and Prince, 2009)), tissue membership attributes (requiring task-specific segmentation) (e.g., (Shen and Davatzikos, 2002; Xing et al., 2008; Zacharaki et al., 2008; Ellingsen and Prince, 2009)), wavelet-based attributes (e.g., (Xue et al., 2004)), local frequency attributes (e.g., (Liu et al., 2002; Jian et al., 2005)), local intensity histogram attributes (e.g., (Shen, 1997; Yang et al., 2008)), geodesic intensity histogram attributes (e.g., (Ling and Jacobs, 2005; Li et al., 2009a)), tensor orientation attributes (specifically for diffusion tensor imaging registration) (e.g., (Verma and Davatzikos, 2004; Munoz-Moreno et al., 2009; Yap et al., 2009)), and curvature attributes (specifically for surface matching) (e.g., (Shen et al., 2001; Liu et al., 2004; Zhan et al., 2007; Ou et al., 2009a)). These studies showed improved registration accuracies as a result of more distinctive characterizations of voxels. However, as most of these attributes are pre-conditioned on segmentation, tissue labeling, or edge detection, and are tailored to some specific image content or modalities, there is still a need to find a common set of attributes that are generally applicable for images containing various organs. In addition, there is also a need to automatically, rather than heuristically, select the optimal attribute components.
To meet the above needs, DRAMMS extracts a set of multi-scale and multi-orientation Gabor attributes at each voxel by convolving the images with a set of Gabor filter banks. The use of Gabor attributes in DRAMMS is mainly motivated by the following three properties of Gabor filter banks:

1) **General applicability and successful application in numerous tasks.** Almost all medical images have texture information, at some scale and orientation, reflecting the underlying geometric and anatomical characteristics. The texture information can be effectively captured by Gabor attributes, as demonstrated in a variety of studies, including texture segmentation (e.g., (Jain and Farrokhnia, 1991)), image retrieval (e.g., (Manjunath and Ma, 1996)), cancer detection (e.g., (Zhang and Liu, 2004)) and tissue differentiation (e.g., (Zhan and Shen, 2006; Xue et al., 2009)). Recently, Gabor attributes have also been successfully used in (Liu et al., 2002; Verma and Davatzikos, 2004; Elbakary and Sundareshan, 2005) to register medical images involving different organs (brain, heart). This promises the use of Gabor attributes for a variety of image registration tasks. They have also pointed out some drawbacks related to the use of Gabor attributes, such as the high computational cost and the need for an appropriate choice of attribute components due to the redundancy of the conveyed information. DRAMMS addresses both drawbacks by developing a module for selecting the optimal Gabor components, which is detailed later in Section 3.2.2.2.

2) **Suitability for single- and multi-modality registration tasks.** This is mainly because of the multi-scale property in Gabor filter banks. Specifically, the high-frequency Gabor filters serve as edge detectors. The detected edge information is, to some extent, independent of the underlying intensity distributions, and is therefore suitable for multi-modality registration tasks. This holds true even when intensity distributions in the two images no longer follow a consistent relationship, in which case mutual-information (Wells et al., 1996; Maes et al., 1997) based methods are challenged (Liu et al., 2002). On the other hand, the low-frequency Gabor filters serve as local image smoothers. The corresponding attribute images are analogous to images at coarse resolutions, which will largely help prevent the cost function from
Figure 12: Multi-scale and multi-orientation Gabor attribute images extracted from two typical human brain images. To save space, shown here are only representative attributes from the real part of three orientations ($\pi/2$, $\pi/3$ and $\pi$) in the $x$-$y$ plane (a full set of Gabor attributes is mathematically described in Eq. 3.5).

being trapped at the local minima. As a demonstration, Fig. 12 shows a typical set of high- and low-frequency Gabor attribute images for two typical human brain images from different modalities. In this case, the edge information extracted by high-frequency Gabor filters (the first row in the figure) provides the basis for the multi-modality registration.

3) Multi-scale and multi-orientation nature. As pointed out in [Kadir and Brady, 2001], scale and orientation are closely related to the distinctiveness of attributes. The multi-scale and multi-orientation attributes are more likely to render each voxel distinctively identifiable, therefore reducing ambiguities in attribute-based voxel matching. For instance, compared with traditional edge detectors (e.g., Canny, Sobel), the multi-orientation Gabor attributes will easily differentiate edges in horizontal and vertical directions (see Fig. 12).
Direct computation of 3D Gabor attributes is usually time consuming. To reduce computational cost, the approximation strategy in (Manjunath and Ma, 1996; Zhan and Shen, 2006) is adopted. Specifically, the 3D Gabor attributes at each voxel are approximated by two groups of 2D Gabor filter banks in two orthogonal planes ($x$-$y$ and $y$-$z$ planes). Mathematically, the two groups of 2D Gabor filter banks in two orthogonal planes are

\[
g_{m,n}(x,y) = a^{-m}g(a^{-m}x',a^{-m}y') \quad \text{in } x$-$y \text{ plane},
\]

\[
h_{m,n}(y,z) = a^{-m}h(a^{-m}y',a^{-m}z') \quad \text{in } y$-$z \text{ plane}; \quad (3.2)
\]

where $a$ is a scalar for re-sampling images to different scales ($a$ is usually set to 2). $m = 1, 2, \ldots, M$ is the index for scales, with $M$ being the total number of scales. As $m$ varies, the scaling factor $a^m$ leads to different sizes of the neighborhood from which the multi-scale Gabor attributes are extracted. $x' = x \cos \left(\frac{n\pi}{N}\right) + y \sin \left(\frac{n\pi}{N}\right)$, $y' = -x \sin \left(\frac{n\pi}{N}\right) + y \cos \left(\frac{n\pi}{N}\right)$, $y'_h = y \cos \left(\frac{n\pi}{N}\right) + z \sin \left(\frac{n\pi}{N}\right)$ and $z'_h = -y \sin \left(\frac{n\pi}{N}\right) + z \cos \left(\frac{n\pi}{N}\right)$ are rotated coordinates, where $n = 1, 2, \ldots, N$ is the index for orientations, with $N$ being the total number of orientations. As $n$ varies, Gabor envelopes rotate in $x$-$y$ and $y$-$z$ planes, leading to multi-orientation Gabor attributes.

\[
g(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp \left[-\frac{1}{2} \left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right) + j2\pi f_x \right]; \quad (3.3)
\]

\[
h(y, z) = \frac{1}{2\pi\sigma_y\sigma_z} \exp \left[-\frac{1}{2} \left(\frac{y^2}{\sigma_y^2} + \frac{z^2}{\sigma_z^2}\right) + j2\pi f_y \right]; \quad (3.4)
\]

$g(x, y)$ and $h(y, z)$ in the equations above are "mother Gabor filters", which are complex-valued functions in the spatial domain. They are each obtained by modulating a Gaussian envelope with a complex exponential. Intuitively, they are elliptically-shaped covers (Gaussian envelopes) that specify the neighborhood around each voxel, from which the Gabor attributes are extracted. Mathematically, the mother Gabor filters $g(x, y)$ and $h(y, z)$ are expressed as
where $\sigma_x$, $\sigma_y$ and $\sigma_z$ are the standard deviations of the Gaussian envelope in the spatial domain; $f_x$ and $f_y$ are modulating (shifting) factors in the frequency domain (often known as “central frequencies”) (Kamarainen, 2003).

As a result of attribute extraction, each voxel $x = (x, y, z)$ is characterized by a Gabor attribute vector $\tilde{A}_i(x)$ with dimension $D = M \times N \times 4$, as denoted in Eq. 3.5. Here, the factor 4 arises because that there are two parts (real and imaginary) of the Gabor responses in two orthogonal planes.

\[
\tilde{A}_i(x) = [\text{Real}((I_i * g_{m,n})(x,y)), \text{Imaginary}((I_i * g_{m,n})(x,y)),
\text{Real}((I_i * h_{m,n})(y,z)), \text{Imaginary}((I_i * h_{m,n})(y,z))]_{m=1,2,...,M,n=1,2,...,N}.
\]

We call this attribute vector $\tilde{A}(\cdot)$ a "full" Gabor attribute vector, in contrast to the "optimal" attribute vector $A^*(\cdot)$ (the selection of the optimal components is explained in Section 3.2.2.2). The numbers of their dimensions are denoted as $D$ and $d$, respectively, whereas $D$ is determined by the number of scales and orientations in the extraction of Gabor attributes ($D = M \times N \times 4$), and $d$ is determined automatically during the subsequent attribute selection process.

Note that in Eq. 3.5 we have taken the magnitude (absolute value) of the Gabor response in the real and imaginary parts. Therefore, the edge information extracted by high-frequency Gabor filters will remain relatively stable even when image contrast changes.

**Role of Full Gabor Attributes.** Before selecting the optimal attributes and using the extracted attribute vectors for cross-image matching, we need to make sure that the extracted full attribute vector $\tilde{A}_i(x)$ in Eq. 3.5 is able to characterize voxels distinctively at least within the same image. By distinctive characterization, we mean that the attributes should be able to differentiate a point from all other points in the image. In Fig. 13, two
special points (labeled by red and blue crosses in sub-figure (a)) and two ordinary points (labeled by orange and purple crosses in sub-figure (a)) are examined in a typical brain MR image. Similarity maps (b,c,d,e) are calculated between each of these examined voxels and all other voxels in the same image. The similarity is calculated based on the full Gabor attribute vectors. Its definition can be found in Eq. 3.6. From these similarity maps it can be observed that, the example special voxels and ordinary voxels are all localized with fairly low levels of ambiguities by the full Gabor attributes $\tilde{A}_i(x)$.

Figure 13: The role of full Gabor attributes in characterizing voxels relatively distinctively within an image. Similarity maps (b-e) are created by calculating attribute-wise similarity between a single point (depicted by either red, blue, orange or purple crosses in (a)) with all points in the same image. Compared with the intensity attribute, the full Gabor attribute vector described in Eq. 3.5 leads to far smaller number of voxels that are similar to the points being examined.
3.2.2.2. Attribute Selection

A disadvantage of Gabor attributes is the redundancy among attributes, which is caused by the non-orthogonality among Gabor filters at different scales and orientations. This redundancy not only increases computational cost, but more importantly, it may reduce the distinctiveness of attribute representation, causing ambiguities in the attribute matching (Manjunath and Ma 1996; Kamarainen, 2003). A learning-based method is therefore designed to select the optimal components.

The main idea of attribute selection is the following: if provided with some pairs of corresponding voxels from the two images, we can treat them as training samples and employ a machine learning method to select the optimal attribute components, such that their similarity and matching reliability are preserved or preferably increased.

Two issues are essential for formulating this idea: (1) the selection of training voxel pairs — in the context of general-purpose registration algorithm like ours, they should be selected from the images to be registered. The selection should be fully automatic without an assumption of any a priori knowledge. Moreover, representative voxel pairs should be those that have high similarity and matching reliability. (2) The selection of optimal attributes — the criterion is to preserve or increase the similarity and matching reliability between the training voxel pairs.

Both issues rely on the quantifications of the similarity and matching reliability between two voxels. The similarity, $\text{sim}(\mathbf{p}, \mathbf{q})$, between a pair of voxels ($\mathbf{p} \in \Omega_1, \mathbf{q} \in \Omega_2$) is defined based on their attribute vectors,

$$\text{sim}(\tilde{\mathbf{A}}_1(\mathbf{p}), \tilde{\mathbf{A}}_2(\mathbf{q})) = \frac{1}{1 + \frac{1}{D_1^2} \| \tilde{\mathbf{A}}_1(\mathbf{p}) - \tilde{\mathbf{A}}_2(\mathbf{q}) \|^2} \in [0, 1].$$  \hspace{1cm} (3.6)

A smaller Euclidean distance between the attribute vectors of two voxels indicates a higher similarity between them. Note that the squared Euclidean distance between two attribute
vectors is normalized by the dimensionality $D$, so as to make the similarity calculated from different number of attributes comparable. The matching reliability is represented by the mutual-saliency weight $ms(\cdot, \cdot)$, which is elaborated later in section 3.2.3. With these quantifications, we can present below the two steps to select the optimal Gabor attribute components.

**Step 1: Selecting Training Voxel Pairs.** To encourage the training voxel pairs to represent different anatomical regions, they should be scattered in the image domain and far away from each other. Accordingly, DRAMMS regularly partitions the subject image $I_1$ into $J$ regions $\Omega_1^{(j)} (j = 1, 2, \ldots, J)$ and selects from each region a voxel pair $(p_j^*, q_j^* \in \Omega_1^{(j)} \subset \Omega_1, q_j^* \in \Omega_2)$ that has the highest matching similarity and matching reliability (measured by mutual-saliency), under the full Gabor attributes. Mathematically,

$$
(p_j^*, q_j^*) = \arg \max_{p \in \Omega_1^{(j)}, q \in \Omega_2} \left[ \frac{\text{sim} (\tilde{A}_1(p), \tilde{A}_2(q)) \times ms(p, q)}{\text{Similarity} \times \text{Mutual-Saliency}} \right] \quad (3.7)
$$

Note that the full Gabor attribute vector $\tilde{A}(\cdot)$ is used at this step, since no selection has been conducted until this stage. The mutual-saliency measure $ms(p, q)$ is explained in a later Section 3.2.3; for now, the bottom line is that it reflects the matching reliability between $p \in \Omega_1$ and $q \in \Omega_2$.

Fig. 14 illustrates the selection of training voxel pairs. Here, the regular partitioning is used instead of a more complicated organ/tissue segmentation, in order to keep DRAMMS as a general-purpose registration method without assumptions on more complicated image segmentation. Note also that the template image $I_2$ is not partitioned because at this stage, no transformation is performed and no corresponding regions should be assumed.

**Step 2: Selecting Optimal Attributes.** Once training voxel pairs $(p_j^*, q_j^*)_{j=1}^J$ have been determined, DRAMMS selects a subset of optimal attribute components $A^*(\cdot)$ out of the full Gabor attribute vector $\tilde{A}(\cdot)$, such that they maximize the overall similarity and matching
reliability on these selected training voxel pairs. Mathematically,

$$\{A_1^*, A_2^*\} = \arg \max_A \sum_{j=1}^J \left[ \frac{\text{sim}(A_1(p^*_j), A_2(q^*_j)) \times \text{ms}(p^*_j, q^*_j)}{\text{Similarity} \times \text{Mutual-Saliency}} \right]$$  \hspace{1cm} (3.8)

The same in Eq. 3.7, $\text{ms}(\cdot, \cdot)$ is the mutual-saliency metric that reflects the matching reliability, which is explained in the subsequent Section 3.2.3.

Actually, the objective functions in Eq. 3.7 and Eq. 3.8 are almost the same, both containing $[\text{sim}(\cdot, \cdot) \times \text{ms}(\cdot, \cdot)]$. The difference is that, we are optimizing over all possible voxel pairs in Eq. 3.7 holding the full attribute vector $\tilde{A}(\cdot)$ unchanged; whereas we are optimizing over all possible attribute subsets in Eq. 3.8 holding the training voxel pairs $(p^*_j, q^*_j)_{j=1}^J$ unchanged.

To implement Eq. 3.8, DRAMMS adopts an iterative backward elimination (BE) and forward inclusion (FI) strategy for attribute selection. Such a strategy is commonly used for attribute/variable selection in the machine learning community (e.g., [Guyon and Elisseeff, 2003; Fan et al., 2007]). The iterative process is specified as follows. Starting from a set of $D$ components in the full Gabor attributes $\tilde{A}(\cdot)$, each time we eliminate one Gabor
component, such that the objective function \([\text{sim}(\cdot, \cdot) \times m_{\text{sim}}(\cdot, \cdot)]\) increases by the largest amount than eliminating any other component. This is known as backward elimination. We continue the elimination process, eliminating one component at a time, until no component can be eliminated to further increase the value of the objective function. This ends one round of backward elimination. Then, we start to include the previously-eliminated Gabor components, one at a time, so that the cost function \([\text{sim}(\cdot, \cdot) \times m_{\text{sim}}(\cdot, \cdot)]\) increases by the largest amount than including any other component. This is known as forward inclusion. We keep on this inclusion process, including one component at a time, until no component can be included to further increase the value of the objective function. This ends one round of forward inclusion. Finally, we iterate between backward elimination and forward inclusion and monotonically increase the objective function, until no attribute components can be eliminated or included, which we call convergence. Convergence is guaranteed as the number of attributes is bounded below and above. Upon convergence, the objective function \([\text{sim}(\cdot, \cdot) \times m_{\text{sim}}(\cdot, \cdot)]\) is maximized and the remaining attributes are the optimal set of attributes we will eventually use for attribute matching, denoted as \(A^*(\cdot)\). Fig. 15 shows a typical attribute selection process for two cardiac images being registered. In this case, the objective function keeps increasing during one round of the BE process (where 39 attributes are eliminated from the full attribute set), one round of the FI process (where 9 previously-eliminated attributes are included back to the remaining attribute set), and one more round of BE processes (where 4 attributes are eliminated from the attribute set), until it reaches the maximum, and then the optimal attribute set has been selected.

**Role of Optimal Gabor Attributes.** Fig. 16 compares the intensity attribute, gray-level-cooccurrence-matrix (GLCM) texture attributes (Haralick, 1979), full Gabor attributes, and the optimized Gabor attributes, in terms of matching ambiguities caused by different attributes. Similarity maps are generated by these attributes between a special/ordinary point in the subject brain/cardiac image and all points in the template image. The similarity value between two voxels is calculated according to Eq. 3.6. The optimal Gabor attributes lead to highly distinctive attribute characterizations, with fewer candidates from
Figure 15: A typical scenario for attribute selection. (a) The value of the objective function $[sim(\cdot, \cdot) \times ms(\cdot, \cdot)]$ with regard to the iteration index. (b) The number of attributes selected with regard to the iteration index. Starting from a full set of attributes, iterative backward elimination (BE) and forward inclusion (FI) processes are employed to select the optimal attributes. The objective function increases monotonically until no other attribute could be eliminated or included to increase it further. When the cost function stops increasing, the corresponding attributes are the ones selected as the optimal set.
Figure 16: The role of the attribute selection in reducing matching ambiguities, as illustrated on special voxels (red crosses) and ordinary voxels (blue crosses) in brain and cardiac images of different individuals. Similarity maps are generated between a voxel (red or blue) in the subject image and all voxels in the template image. "GLCM", gray-level co-occurrence matrix (Haralick, 1979), is another commonly used texture attribute descriptor.

the other image to match up with a query voxel, compared with using other attribute characterizations. This holds true for a special point (depicted by red crosses) and for an ordinary point (depicted by blue crosses). Since the similarity map for a given voxel now looks more like a "delta" function in the neighborhood of this voxel, the optimal Gabor attributes offer great promise in reducing the risk of being trapped at local minima.
3.2.3. Mutual-Saliency Weighting to Modulate Registration

The above Section 3.2.2 describes extracting a full set of Gabor attributes (3.2.2.1) and selecting the optimal Gabor attributes (3.2.2.2) to characterize each voxel distinctively. That finishes the first component of DRAMMS, namely, attribute matching (AM). This section elaborates the second component of DRAMMS, namely, mutual-saliency (MS) weighting, to better modulate the registration process.

As addressed in the introduction of this chapter, an ideal optimization process should utilize all voxels but assign a continuously-valued weight to each voxel, based on the capability of this voxel to establish reliable correspondences across images. The idea of quantifying matching reliability was perhaps first investigated in (Anandan, 1989; Duncan et al., 1991) and their follow-up studies (McEachen and Duncan, 1997; Shi et al., 2000). They developed a "confidence" quantity for motion tracking on surface points, which measures whether the matching of two surface points is reliable in the neighborhood, based on the similarity defined on their curvatures. Other works explicitly segment the regions of interest (ROI) in a joint segmentation and registration framework (Yezzi et al., 2001; Wyatt and Nobel, 2003; Chen et al., 2005; Pohl et al., 2006; Wang et al., 2006; Xue et al., 2008; Ou et al., 2009a; Periaswamy and Farid, 2006) and assign higher weights in segmented ROIs. Since both approaches depend on explicit detection or segmentation of certain feature voxels/regions, they are not generally applicable to diverse tasks. For an extension, we aim to develop a quantity that measures matching reliability for each and every single voxel in the image; also, the quantification method should not rely on any prerequisite detection or segmentation.

Recent work (Huang et al., 2004; Bond and Brady, 2005; Yang et al., 2006; Wu et al., 2006; Mahapatra and Sun, 2008) assumed that more salient regions could establish more reliable correspondences and hence should be assigned with higher weights. They reported improved registration accuracies in a number of registration tasks. However, this intuitive assumption does not always hold true, because regions that are salient in one image are not
necessarily salient in the other image, or more importantly, do not necessarily have reliable correspondences between images. To visualize this problem, a simple counter-example is shown in Fig. 9: the boundary of the lesion is salient in the diseased brain image, but it does not have a correspondence in the normal brain image, so assigning higher weights to it will harm the registration process. In other words, (single) saliency in one image does not necessarily indicate matching reliability between two images.

To remedy this problem, (Luan et al., 2008) extended the single saliency criterion into a double (joint) saliency criterion. In their work, higher weights are assigned to a pair of voxels if they are respectively salient in two input images. However, two nearby points may belong to completely different anatomical structures even though they are both salient in their own images. For instance, also in Fig. 9, the boundary of the lesion is salient in the subject image, whereas another voxel at the same spatial location may be coincidently salient in the other image, though they belong to different tissue types. The dilemma in the double (joint) saliency metric is actually similar to the dilemma in the single saliency metric — the saliency is individually measured in one image, so it does not necessarily indicate the matching reliability across images.

To directly measure matching reliability across two images, DRAMMS extends the concepts of single saliency or double saliency into the concept of mutual-saliency. Generally speaking, the matching is reliable if the two points are similar to each other, and are not similar to anything else in the neighborhood. In this case, the similarity map between a point \( x \) in one image and all the points in the neighborhood in the other image should exhibit a "delta" shaped function, with the pulse located right at the corresponding point \( h(x) \), as shown in Fig. 17. Therefore, to calculate the mutual-saliency value \( ms(x, h(x)) \), we need to quantitatively check the existence and the height of a delta function in the similarity map generated between voxel \( x \) and all voxels in the neighborhood of \( h(x) \). This is the idea of the mutual-saliency metric.

This idea is formulated in Eq. 3.9 and the associated Fig. 18.
Figure 17: The idea of the mutual-saliency measure. The matching between a pair of voxels \( x \) and \( h(x) \) is reliable if they are similar to each other and not similar to anything else in the neighborhood. Therefore, a delta function in the similarity map indicates an reliable matching, and hence a high mutual-saliency value.

\[
ms(x, h(x)) = \frac{\text{MEAN}_{v \in CN(h(x))} \left[ \text{sim}(A_1(x), A_2(v)) \right]}{\text{MEAN}_{v \in PN(h(x))} \left[ \text{sim}(A_1(x), A_2(v)) \right]}
\]

(3.9)

where \( \text{sim}(\cdot, \cdot) \), the attribute-based similarity between two voxels, is defined in the same

Figure 18: Explanation of the mutual-saliency definition in Eq. 3.9. Different colors encode different layers of neighborhoods. A high average similarity in the core neighborhood (CN) and a low average similarity in the peripheral neighborhood (PN) indicate a reliable matching and hence a high mutual-saliency value. This figure is better viewed in the color version.
way as in Eq. 3.6. Note that voxels in between the core and peripheral neighborhoods of $h(x)$ are not included in the calculation, because there is typically a smooth transition from high to low similarities, especially for coarse-scale Gabor attributes. For the same reason, the radii of these neighborhoods are adaptive to the scale in which Gabor attributes are extracted. In particular, in accordance with the Gabor parameter that we will discuss in Section 3.5.2, the radii of the core, transitional and peripheral neighborhoods are 2, 5, 8 voxels, respectively, for a typical isotropic 3D brain image.

In the implementation, the mutual-saliency map is updated each time the transformation $h$ is updated, such as in an expectation-maximization (EM) strategy, since mutual-saliency $ms(\cdot, \cdot)$ and transformation $h$ are affecting each other by Eqs. 3.9 and 3.1.

**Role of Mutual-Saliency Maps.** The mutual-saliency map effectively identifies outlier regions, whose correspondences do not present in the other image, and assigns low weights to them. This helps reduce their negative impact in the optimization process. In Fig. 19, motivated by our work on matching histological sections with MRI, we have simulated cross-shaped tears (cuts) in the subject image, which are typical situations when sections are stitched together. The mutual-saliency map assigns low weights to the tears/cuts, therefore reducing their negative impact towards registration. On the contrary, registration without using the mutual-saliency map tends to fill in the tears by over-aggressively pulling other regions. This causes artificial results as pointed out by the arrows in Fig. 19(c). Besides, by using the mutual-saliency weighting framework, the deformation field in Fig. 19(g) is smoother and more realistic than the one without using the mutual-saliency map framework, which is shown in Fig. 19(d).
Figure 19: The role of the mutual-saliency map in accounting for outlier regions and the consequent partial loss of correspondence. (a) The subject image, with simulated deformation and tears from (b), the template image. (c) The registered images without using the mutual saliency map; (d) The deformation field associated with (c); (e) The registered image using the mutual-saliency map; (f) The mutual-saliency map associated with (e); (g) The deformation field associated with (e). With red points marking spatial locations having exactly the same coordinates in all the sub-figures, we show that the upper-left corner of the gray patch (pointed by yellow arrow in (c)) are initially corresponding between the source (a) and the target (b) image, and this correspondence is preserved in the registered image (e) when the mutual-saliency map is used. On the other hand, the correspondence is destroyed in the registered image (c) when the mutual-saliency map is not used. Since this corner point is right next to the image cut, the preservation of correspondence demonstrates the decreased negative impact caused by missing data when the mutual-saliency map is used.
3.3. Numerical Implementation

A complete registration framework usually consists of three parts: 1) a similarity measure, which defines the criterion to align two images; 2) a deformation model, which defines the mechanism to transform one image to the other; and 3) an optimization strategy, which is used to calculate the best parameters of the deformation model such that the similarity between the two images is maximized.

The attribute matching and mutual-saliency weighting components together in DRAMMS (Eq. 3.1) are essentially a new definition of image similarity measure, namely the first part of a registration framework. It could be readily combined with any deformation model and any optimization strategy to form a complete registration framework. For the sake of completeness, we have specified free form deformation (FFD) as the deformation model and the discrete optimization as the optimization strategy in the following two sub-sections.

3.3.1. Choice of Deformation Model

A wide variety of voxel-wise deformation models can be selected, such as demons (Thirion, 1998; Vercauteren et al., 2009), fluid flow model (Christensen et al., 1994; DAgostino et al., 2003), and free form deformation (FFD) model (Rueckert et al., 1999, 2006). In our implementation, the diffeomorphic FFD (Rueckert et al., 2006) is chosen, because of its flexibility to handle a smooth and diffeomorphic deformation field, and its popularity among the deformable registration community. In the following experiments, the distance between the control points is chosen at 7 voxels in the x-y directions and 2 voxels in the z direction. The deformation is diffeomorphic by constraining the displacement at each control point (node) in the FFD model not to exceed 40% of the spacing between two adjacent control points (nodes). Previous work (Choi and Lee, 2000) has shown that this 40% hard constraint
will guarantee diffeomorphism of deformation fields. Note that, this 40% hard constraint is the maximum displacement in each iteration of each resolution. Therefore, composition of transformations in multiple iterations and multiple resolutions can still cope with large deformations while maintaining diffeomorphism.

### 3.3.2. Choice of Optimization Strategy

To optimize the parameters for the FFD deformation model, a numerical optimization strategy is needed. A large pool of optimization strategies can be used, including gradient descent (Cauchy 1847), Newton’s method (Avriel 2003), Powell’s method (Powell 1964), and the discrete optimization strategy (Komodakis et al. 2008). In our implementation, we have chosen the discrete optimization strategy, a state-of-the-art optimization strategy known for computational efficiency and robustness with regard to local optima (Komodakis et al. 2008; Glocker et al. 2008). Loosely speaking, the discrete optimization strategy densely samples the displacement space in a discrete manner, and seeks a group of displacement vectors on the FFD control points (nodes) in a combinatorial fashion, so that they collectively minimize the matching cost. Table 1 compares the computational times between the gradient descent optimization strategy and the discrete optimization strategy. The speedup by using the discrete optimization strategy is not difficult to be observed. Since the choice of a numerical optimization strategy is not the primary contributions of the proposed registration algorithm, we do not include any further comparisons. More comparisons can be found though, in Glocker et al. 2008.

For the sake of completeness of this section, the discrete optimization strategy is briefly described below. Interested readers are referred to Komodakis et al. 2008; Glocker et al. 2008 for more detail.

Let $\Phi = \{\phi\}$ be the entire set (grid) of control points (nodes) $\phi$, then the energy in Eq. 3.1
Table 1: Demonstration of high computational efficiency of the discrete optimization (DisOpt) strategy compared with the gradient descent (GradDes) strategy. Registration accuracy from the two optimization strategies is almost equivalent, so it is not listed in this table. Computational times are recorded on registering two (inter-subject) brain images, two (inter-subject) cardiac images and two (multi-modality) prostate images, for attribute matching (AM) both with and without mutual-saliency (MS) weighting. Unit: minutes.

<table>
<thead>
<tr>
<th></th>
<th>FFD + GradDes</th>
<th>FFD + DisOpt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AM w/ MS</td>
<td>AM w/o MS</td>
</tr>
<tr>
<td>Brain images</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(256 × 256 × 171)</td>
<td>534.67</td>
<td>268.82</td>
</tr>
<tr>
<td>Prostate images</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(256 × 256 × 34)</td>
<td>245.46</td>
<td>104.77</td>
</tr>
<tr>
<td>Cardiac images</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(256 × 256 × 20)</td>
<td>181.54</td>
<td>88.77</td>
</tr>
</tbody>
</table>

can be rewritten after being projected onto the nodes as:

\[
E(h) = \frac{1}{|\Phi|} \sum_{\phi \in \Phi} \left( \int_{x \in \Omega_1} \eta^{-1}(\|x - \phi\|) \cdot ms(x, h(x)) \cdot \|A^*_1(x) - A^*_2(h(x))\|^2 dx + \lambda R(\nabla \Phi h_\phi) \right)
\]

(3.10)

where \( \eta^{-1}(\cdot) \) is an inverse mapping function that maps the influence from a node \( \phi \in \Phi \) to every ordinary point \( x \); \( h_\phi \) is the translation that is applied to the node \( \phi \) of the grid \( \Phi \). Compared to (Glocker et al., 2008), here the inverse weighting function \( \eta^{-1}(\cdot) \) acts only as an apodizer function, which does not give weights to the voxels with respect to their distance from the control points. The only weights given are the ones due to the mutual saliency value \( ms(\cdot, \cdot) \).

Following (Glocker et al., 2008), the registration problem is cast as a multi-labelling one and the theory of Markov Random Fields (MRF) is used to formulate it mathematically. The solution space is discretized by sampling along the \( x, y \) and \( z \) directions as well as along their diagonals, resulting in a set \( L \) of \((18 \times n + 1)\) labels, where \( n \) is the number of sampling labels (discretized displacement) along each direction. What we search now is which label to be attributed to each node of the deformation grid, i.e., which displacement to be applied to each node. The energy in Eq. 3.10 can be approximated by a MRF energy...
whose general form is the following:

$$E_{MRF}(I) = \sum_{\phi \in \Phi} \left( V_{\phi}(l_{\phi}) + \sum_{\psi \in N(\phi)} V_{\phi\psi}(l_{\phi}, l_{\psi}) \right)$$  \hspace{1cm} (3.11)$$

where \( I \) is the labeling (discretized displacement) that we search, \( l_{\phi}, l_{\psi} \in L \) and \( V_{\phi}, V_{\phi\psi} \) are the unary and the pair-wise potentials respectively. The unary and the pair-wise potentials encode the data and the regularization term of the energy in Eq. (3.10). Specifically, for the case of the data term:

$$V_{\phi}(l_{\phi}) = \int_{x \in \Omega_1} \eta^{-1}(\|x - \phi\|) \cdot ms(x, h(x)) \cdot \|A^*_1(x) - A^*_2(h(x))\|^2 dx$$  \hspace{1cm} (3.12)$$

The regularization will be defined in the label domain as a simple vector difference, thus:

$$V_{\phi\psi}(l_{\phi}, l_{\psi}) = \|h_{\phi} - h_{\psi}\|.$$  \hspace{1cm} (3.13)$$

This encourages nearby control points (nodes) to have consistent geometric displacements.

In this discrete optimization framework, the number of labels (discretized displacement) in the search space is the key factor to balance between the registration accuracy and computational efficiency. More labels correspond to denser sampling of the possible displacements, therefore a higher accuracy but lower computational efficiency. A good trade-off is obtained by a multi-resolution approach, where greater deformations (coarsely-distributed labels in a large search space) are captured in the beginning and the solution will be refined in every step by considering smaller deformations (densely-distributed labels in a small search space). This balances the computational cost and the registration accuracy.

A list of parameters in this implementation is provided as follows. The regularization parameter \( \lambda \) is set to 0.1 for the discrete optimization case in Eq. (3.10). The \( \eta^{-1}(\cdot) \) indicator function is defined in such a way that all the voxels belonging to a 26-neighborhood centered to a node contribute to the energy projected to it. \( n = 4 \) labels are sampled per direction.
The distance between the control points is chosen at 7 voxels in the $x - y$ directions and 2 voxels in the $z$ direction. All experiments are operated in C code on a Intel® Xeon® Processor (2.80 GHz, 4M Cache, 800 MHz FSB) with LINUX operation system, and the computational times using the discrete optimization strategy in several typical scenarios can be found in the rightmost two columns in Table 1.

### 3.4. Results

This sections validates DRAMMS in a number of registration tasks. It serves an initial validation section to briefly show a) generality b) accuracy and c) robustness to outlier regions. In Chapter 4, the issue of validating DRAMMS will be re-visited at greater depth, in a quantitative fashion, compared with 12 state-of-the-art registration methods, in more registration tasks (brain (skull-striped, raw, lesion-bearing, tumor-recurring), cardiac, breast), and in large-scale public and in-house datasets.

In this section, DRAMMS is tested in various registration tasks containing different image modalities and organs. The results are compared with those obtained from mutual information (MI)-based FFD, a commonly-used deformable registration method. A public software package MIPAV (Medical Image Processing, Analysis and Visualization) ([McAuliffe et al., 2001](https://www.nih.gov/)) available from NIH is used for MI-based FFD registration. The use of MIPAV is because of its user-friendly cross-platform interface, its computational efficiency and its popularity within the society. Specifically, the MI-based FFD in MIPAV is implemented in multi-resolution and by Powell’s optimization strategy.
3.4.1. Simulated Images

Registration results on a set of simulated images have already been shown in Fig. 19. We have simulated regions of missing correspondences by generating a cut in the subject image as shown in Fig. 19(a). DRAMMS largely reduces the negative impact caused by outlier regions, and therefore, generates anatomically realistic results than that from MI-based FFD.

3.4.2. Inter-Subject Registration.

Our second set of experiments is performed on registration of brain images and registration of cardiac images from different individuals (the ones shown in the left column of Fig. 16). The images being registered are of the same modality. Therefore, the registration results can be quantitatively compared in terms of mean squared difference (MSD) and correlation coefficient (CC) between the registered and the template images — a high registration accuracy should correspond to a decreased MSD value and an increased CC value. This evaluation criterion has also been used in Rueckert et al. (1999). In this experiment, we recorded registration accuracy obtained by intensities, by Gabor attributes, and by the optimal Gabor attributes, both with and without mutual-saliency weighting. Overall, Fig. 20 shows that, each of DRAMMS’ components provides additive improvement of registration accuracy over MI-based FFD. The largest improvement over MI-based method occurs at replacing the one-dimensional intensity attribute with high-dimensional Gabor attributes.

3.4.3. Brain Atlas Construction.

Our third set of experiments consists of atlas construction of human brain images. Fig. 21 shows 30 randomly selected brain MR images used for constructing an atlas. Each brain
Figure 20: The quantitative evaluation of registration accuracies of inter-subject brain and cardiac images, in terms of MSD and CC between registered and template images. The various parts of DRAMMS – feature extraction, feature selection, mutual-saliency weighting – are added into the experiment sequentially. As a result, this figure shows that each of DRAMMS’ components provides additive improvement over MI-based FFD.

image is registered (in 3D) to the template (outlined by the red box). Then the warped images are averaged to form an atlas. The constructed atlases are shown in Fig. 22 from there DRAMMS is observed to result in sharper average, indicative of a higher registration accuracy on average. Differences between the constructed atlases and the template image are further visualized, as shown in the top row in Fig. 23. Also shown, in the bottom row of Fig. 23 is the voxel-wise standard deviation among all warped images, by affine, MI-based FFD and DRAMMS methods, respectively. The smaller difference with the template in the top row and the smaller standard deviation among all warped images in the bottom row indicate the accuracy and the consistency of DRAMMS method over MI-based FFD. Fig. 24 shows mean squared difference (MSD), correlation coefficient (CC), and errors against expert-defined landmarks, between each warped subject image and the template image, by three registration methods, respectively. We especially emphasize Fig. 24(c), where DRAMMS recovers voxel correspondences closer to expert-defined locations in each single pair-wise case compared with the MI-FFD method.

3.4.4. Multi-modality Registration of the Same Prostate.

Our fourth set of experiments includes some fairly challenging multi-modality registration tasks. Normally, mutual information method is used for multi-modality registration, as long
Figure 21: Thirty randomly selected brain images used for constructing the atlas. The template image is outlined by a red box.
Figure 22: Brain atlases obtained from the 30 randomly-selected images in Fig. (a) The template, the same as the one in the red box in Fig. (b) The atlas by affine registration; (c) Atlas obtained by the MI-based FFD method; (d) The atlas obtained by DRAMMS. The sharper average in (d) indicates a higher registration accuracy for DRAMMS on average.
Figure 23: Further comparisons of affine, MI-based FFD and DRAMMS registration algorithms in the atlas construction. (a1-c1) The difference image ($=\|\text{atlas-template}\|$) obtained from these three registration methods; (a2-c2) The standard deviation among all the warped images, by these registration methods, respectively. The smaller difference from the template image and the smaller standard deviation among all warped images indicate the accuracy and consistency of DRAMMS.
Figure 24: Quantitative comparisons of registration methods for each subject. (a) The mean squared difference (MSD) between the warped image and the template image for all foreground voxels. (b) The correlation coefficient between the warped image and the template image. (c) The errors (in voxels) against manually-defined landmark correspondences.
as intensity distributions from two images follow a consistent relationship. However, when this assumption of a consistent relationship in intensity distributions is violated, mutual information based methods are usually challenged. The registration between histological and MR images shown in Fig. 25 is one example.

For a brief background, histological images are usually taken in the *ex vivo* environment, by physically sectioning an organ into slices and examining the pathology of the slices under a microscopy. Because of the ability to reveal pathologically-authenticated ground truth for lesions and tumor, it is often used as reference to label lesions and tumor in *in vivo* MR images of the same organ, to help future disease detection directly in MRI (Meyer et al., 2006; Dauguet et al., 2007; Zhan et al., 2007; Ou et al., 2009a). Therefore, histology-MRI registration is often needed.

Histology-MRI registration is challenged by the differences in the imaging principles (*ex vivo* versus *in vivo*), image resolutions (µm versus mm level) and image contrasts, which lead to an inconsistent relationship in their intensity distributions. For instance, in Fig. 25 we can observe that dark regions corresponding to dark regions across images (blue circle), white regions corresponding to white regions across images (blue circles), but also dark regions corresponding to white regions (purple circles). In addition, there are some structures that are clearly present in one image but almost invisible in the other. In this case, mutual information methods tend to be challenged, as demonstrated in a number of studies (Pitiot et al., 2006; Meyer et al., 2006; Dauguet et al., 2007). To make things worse, tears/cuts often appear in the histological images, causing missing correspondences. In Fig. 25 crosses of the same color are placed at the same spatial locations in sub-figures (b), the MR image, and (d), the warped histological image, in order to visually reveal whether the anatomical structures have been successfully aligned. Compared with the artificial results obtained by MI-based FFD, DRAMMS leads to a smooth deformation field that better aligns tumors and other complicated structures.
Figure 25: The multi-modality registration between (a) histological and (b) MR images of the same prostate. Crosses of the same color denote the same spatial locations in all images, in order to better visualize the registration accuracy. Circles of the same color roughly denote the corresponding regions, which are used to demonstrate the lack of consistent relationship in their intensity distributions.
3.5. Discussion

In this chapter, a general purpose registration method named DRAMMS is presented. DRAMMS characterizes each voxel by an optimized multi-scale and multi-orientation Gabor attribute vector, which increases the distinctiveness of each voxel and reduces the matching ambiguity. A second feature is that DRAMMS assigns continuously-valued weights for each voxel, based on a mutual-saliency measure that evaluates whether this voxel is capable of establishing reliable correspondences. With the assistance of this mutual-saliency weighting mechanism, image registration relies more on regions that can be effectively identified and matched between images. Therefore, the negative impact caused by missing data and/or partial loss of correspondences is reduced. Experiments in simulated images, cross-subject images and multi-modality images from human brain, heart and prostate have demonstrated the general applicability and registration accuracy of DRAMMS.

This section discusses the various aspects of DRAMMS. Sub-section 3.5.1 discusses how DRAMMS bridges the gap between voxel-wise methods and landmark/feature-based methods. Sub-sections 3.5.2, 3.5.3 and 3.5.4 discuss, respectively, the attribute extraction, attribute selection and mutual-saliency weighting components in DRAMMS. In the end, we discuss the difference between DRAMMS and a closely-related work HAMMER (Shen and Davatzikos, 2002) in Sub-section 3.5.5. The whole chapter is conclude in Sub-section 6.1.4.

3.5.1. DRAMMS as a Bridge between Voxel-wise and Landmark/feature-based Methods

The two components in DRAMMS can be regarded as bridges between the voxel-wise methods and the landmark/feature-based methods in two facets: voxel characterization and voxel utilization. This is listed in Table 2.
Table 2: A List of how DRAMMS bridges the gap between voxel-wise and landmark/feature-based registration methods.

<table>
<thead>
<tr>
<th>Voxel Characterization</th>
<th>Voxel-wise</th>
<th>Landmark/feature-based</th>
<th>DRAMMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voxel Characterization</td>
<td>Intensity</td>
<td>Task-specific attributes</td>
<td>Optimal Gabor attributes</td>
</tr>
<tr>
<td></td>
<td>(generally applicable to various tasks, but not distinctive for voxels)</td>
<td>(distinctive for voxels, but not generally applicable to various tasks)</td>
<td>(generally applicable to various tasks, and distinctive for voxels)</td>
</tr>
<tr>
<td>Spatial Adaptivity</td>
<td>Not adaptive</td>
<td>Binarily adaptive</td>
<td>Continuously Adaptive</td>
</tr>
<tr>
<td>(Voxel Utilization)</td>
<td>(Weight ≡ 1 for all voxels)</td>
<td>(Weight=1 for landmarks/features and 0 otherwise)</td>
<td>(Weighting voxels continuously by mutual-saliency measure)</td>
</tr>
</tbody>
</table>

3.5.2. Discussion on Attribute Extraction

Gabor attributes are used in DRAMMS framework because of the three reasons mentioned in Section 3.2.2.1: general applicability; suitability in single- and multi-modality registration; and the multi-scale and multi-orientation formulation.

When using Gabor filter banks, the number of scales, the number of orientations, and the frequency range to be covered should all be carefully tuned. Ineffectively setting these parameters would unnecessarily introduce a large redundancy among the attributes extracted, increasing the computational burden and decreasing the distinctiveness of attribute description. In this section, we have adopted the parameter settings proposed by (Manjunath and Ma [1996]), which have found success applications in a number of studies such as (Zhan and Shen [2006]). In particular, the number of scales is set at \( M = 4 \) and the number of orientations is set at \( N = 6 \), the highest frequency is set at \( 0.4 \text{mm}^{-1} \), the lowest frequency at \( 0.05 \text{mm}^{-1} \), and the size of the Gaussian envelope depends on the highest frequency by a fixed relationship specified in equation 1 of their paper. Fig. 26 highlights the differences in the consequent similarity maps between adopting their set of parameters (sub-figures b, c, d, e) and adopting a random set of parameters (sub-figures b’, c’, d’, e’).

There are, of course, other texture and geometric attributes that can easily fit into the framework of DRAMMS. Examples include gray-level cooccurrence matrix (GLCM) attributes
Figure 26: Different similarity maps when using different parameter settings in Gabor filter banks for attribute extraction. All similarity maps are calculated between a cross point and all other points in the same image. (b-e) are similarity maps resulted from using Gabor parameters suggested in (Manjunath and Ma, 1996), while counterparts (b’-e’) are resulted from using a random set of Gabor parameters. This figure highlights the importance of adopting the right set of Gabor parameters.

(Haralick, 1979), spin image attributes (Lazebnik et al., 2005), RIFT (rotation-invariant feature transform) attributes (Lazebnik et al., 2005) and fractal attributes (Lopes and Betrouri, 2009). Although Gabor attributes are believed to be a proper choice because of the three merits listed in Section 3.2.2.1 and because of the experimental comparison in Fig. 16, a rigorous comparative study among all of these attributes in the registration context would be of interest. Such a study is a non-trivial task. It requires a large number of images to be registered, requires an objective evaluation of registration accuracies, which is difficult. It also requires, in a mathematical rigorous study, the quantitative analysis of the number and the depth of local minima in the cost function as a result of different attributes or their different combinations. We mark such a study as one of our future research directions.
3.5.3. Discussion on Attribute Selection

Although the overlaps/redundancies among Gabor filter banks can be largely reduced by carefully choosing a set of parameters in the very beginning, the redundancy will almost never vanish completely, simply because of the non-orthogonal nature among Gabor filters. Therefore, the attribute selection module is needed to further reduce the information redundancy and also the computational burden.

The most important part in the selection of attributes is the quantification of matching accuracy and matching reliability, especially the latter. Quantification of the reliability of matching by the mutual-saliency metric enables us to find voxels and attributes components that not only render true correspondences similar, but more importantly, render them reliably similar, meaning similar to each other and not similar to anything else in the neighborhood.

Another nice feature of the attribute selection part is that no \textit{a priori} knowledge is required. It is designed to take no other input except the two input images, thereby keeping DRAMMS a general-purpose method.

Nevertheless, if \textit{a priori} knowledge is given, DRAMMS framework can be readily extended to incorporate it without much effort. For instance, if there are a number of anatomically corresponding voxels given by experts, then we can directly rely on them to select the optimal attributes and simply skip the step of selecting the training voxel pairs, as long as the expert-defined voxel pairs are 1) mutually-salient, 2) representative of anatomic/geometric context in images and 3) spatially distributed. \textit{A priori} knowledge can be also obtained if registration is restricted to a specific type of images (\textit{e.g.} human brain MRI). In this case, a common set of optimal attributes can be learned when registering a large set of such images.
3.5.4. Discussion on Mutual-Saliency Weighting

The mutual-saliency value is automatically derived to reflect whether the matching between a pair of voxels is reliable in the neighborhood. Since the saliency is measured individually in one image and the matching reliability is measured between images, neither single saliency nor double saliency in the existing methods necessarily indicate matching reliability. This is the motivation for developing the mutual-saliency metric, which directly measures the reliability of the matching between two images. Consequently, it can more effectively capture regions/voxels useful for registration and identify regions/voxels having difficulties establishing correspondences.

The mutual-saliency map seems to be most helpful in the presence of outlier regions, also known missing correspondences. A typical example is the image cuts/stitches when registering histological images with MR images (such as shown in Figs. 19 and 25). Another typical example is the lesions/tumors when registering patient images with images of healthy subjects (e.g. Fig. 9). In these cases, existing methods without using mutual-saliency weighting tend to over-aggressively match structures that do not correspond to each other. Although the over-aggressive matching usually leads to smaller intensity residuals (like in Fig. 19(c)), they are not desirable because of the artificial results and the often convoluted deformations. In contrast, the mutual-saliency maps guide registration to match the structures that should have been matched, and leave the structures that have difficulty finding correspondences almost untouched. This leads to more anatomically meaningful results, like in Fig. 19(f). The relative larger intensity residual becomes a right descriptor to encode the missing data.

It should be noted that the mutual-saliency calculation itself is usually computationally costly, because it checks voxel-wise similarities in a neighborhood for each individual voxel, and on top of that, because all these similarities are calculated from high-dimensional attribute vectors. This computational burden can be observed in Table 1 where incorporating the mutual-saliency in registration takes considerably more computational time. Therefore,
for those registration tasks that do not involve missing correspondences, the mutual-saliency component can be dropped, with a noticeable speedup and only a moderate decrease of registration accuracy.

### 3.5.5. Differences from HAMMER algorithms

We refer to the HAMMER registration algorithm ([Shen and Davatzikos 2002](#)) as it is the one closest to DRAMMS. The HAMMER algorithm pioneers the attribute-matching concept, where it extracts geometric-moment-invariant (GMI) attributes, tissue membership attributes and boundary/edge attributes for brain image registration. We point out the following differences between the two methods:

1. **Different applications**: HAMMER is specifically designed for brain image registration, because it is based on attributes from brain segmentation and tissue labeling; whereas DRAMMS is generally applicable to images of various modalities and organs.

2. **Different voxel utilization**: HAMMER is essentially a landmark/feature-based method, which may inherit challenges that are inherent to landmark/feature-based methods; whereas DRAMMS is a voxel-wise method utilizing all imaging data and does not need heuristic thresholding for voxels.

3. **Different ability to handle outlier regions**: compared to HAMMER, DRAMMS is able to handle outlier regions such as brain lesions and histology cuts because of the mutual-saliency mechanism.

4. **Different deformation mechanism**: HAMMER tends to aggressively match brain structures leading to non-diffeomorphic mapping; whereas DRAMMS enforces diffeomorphism in the transformation.
In this chapter, DRAMMS is only compared with other general-purpose methods, and it will be interesting to compare with those specific-purpose methods including HAMMER in future studies, to better establish the registration accuracy of DRAMMS. In Chapter 4, when DRAMMS is extensively compared with 12 state-of-the-art general-purpose registration methods, HAMMER is not included in the comparison. This is because HAMMER is essentially a task-specific algorithm that only applies to skull-stripped brain images. In Chapter 6, we examine the effect of the automatic tissue segmentation, which HAMMER uses, to the accuracy of image registration methods in skull-stripped healthy brain images. It is the first step towards fairly comparing a task-specific method like HAMMER, which relies on tissue segmentation, and all other general-purpose registration methods.

3.5.6. Contributions

In summary, a general-purpose image registration method named "DRAMMS" is presented in this chapter. In DRAMMS, more distinctive characterization of voxels leads to reduced matching ambiguities, and spatially adaptive utilization of imaging data leads to robustness with regard to missing correspondences. Future work, which is outside of the scope of this dissertation, includes comparing different texture/geometric attributes in the framework, better understanding the effect of mutual-saliency in registration scenarios with outlier regions, and more extensively comparing our method with existing ones in various registration tasks.
CHAPTER 4

Validations of DRAMMS

This chapter validates DRAMMS and compares it with 12 state-of-the-art registration methods in various image registration tasks (cross-subject and longitudinal) involving images of various organs (brain, heart and breast). The purpose is to demonstrate three desirable properties of the DRAMMS registration method: (a) generality; (b) accuracy; and (c) robustness to the partial data and missing correspondences. The validations in this chapter lay the foundations for applying DRAMMS to a wide variety of studies, including the five studies mentioned in Chapter 1 that motivate this dissertation.

Each of the following sections validates DRAMMS in one image registration task. The first four sections concern brain image registration tasks. In order of increasing difficulty, we examine: registration of skull-stripped brain images (Section 4.1); registration of raw brain images (Section 4.2); registration of lesion-bearing brain images to images of healthy subjects (Section 4.3); and registration of tumor-recurred brain images to images of healthy subjects (Section 4.4). Then, two other sections validate DRAMMS in other organs: Section 4.5 for cardiac image registration (cross-subject) and Section 4.6 for breast image registration (longitudinal). Finally, Section 4.7 summarizes and concludes this chapter.
4.1. In Registering Skull-stripped Brain MR Images Across Subjects

4.1.1. Introduction

Brain image registration is an extensively studied field. Over the past two decades, a large number of registration methods have been developed (see [Maintz and Viergever 1998; Lester and Arridge 1999; Hill et al. 2001; Zitova and Flusser 2003; Pluim et al. 2003; Crum et al. 2004; Holden 2008] for detailed surveys). A thorough evaluation of many methods in brain registration has drawn increasing interest, yet it remains a demanding task. This is largely due to the lack of ground truth deformations, the difficulty in alternatively defining a comprehensive and unbiased set of evaluation criteria, and the limited computational power to handle registrations in large scale datasets.

Two studies ([West et al. 1997; Hellier et al. 2003]) pioneered evaluation studies by comparing 3 and 6 methods at expert-defined fiducial locations and in cortical areas in cross-subject brain image registration. Recently, many studies ([Yanovsky et al. 2009; Yassa and Stark 2009; Wei 2009]) have emerged. They evaluated up to 6 more recent methods, mostly focusing on specific tasks such as the detection of atrophy and the study of memory.

In a recent paper and perhaps the most thorough evaluation study so far, Klein and colleagues ([Klein et al. 2009]) compared 14 methods (AIR, ANIMAL, ART, Diffeomorphic Demons, fnirt, IRTK, JRD-fluid, ROMEO, SICLE, ANTs, and 4 variations of SPM-embedded methods including DARTEL). They used 8 accuracy indicators based brain regions labeled by human experts to measure the accuracy of registration in 80 brain MR images. They reported ANTs and ART as two most accurate methods among the 14 methods, in cross-subject registration of brain MR images.
This section evaluates DRAMMS against those top-ranked registration methods. We use the same evaluation framework as used in (Klein et al., 2009) for cross-subject registration of brain images. Furthermore, the work presented here has the following features compared with existing validation/evaluation studies:

1. Three recently developed methods (Demons (Vercauteren et al., 2009), DROP (Glocker et al., 2008) and DRAMMS (Ou et al., 2011)) are included because of their reported accuracy, and their representations of recent advances in image registration methodology (similarity metric, optimization strategy);

2. While most studies have only examined deformation accuracy (such as in (Klein et al., 2009)), we thoroughly examine both the accuracy and the aggressiveness of deformations, to better understand their correlation. Deformation aggressiveness is related to the effective degrees of freedom on the deformations, and is explained further in Section 4.1.2.3;

3. While most studies provide only one set of parameters for each method (usually yielding highly conforming deformations), we include both smooth and aggressive parameter settings for top-ranking methods, providing more informed choices under varying needs.

4.1.2. Validation Protocol

This subsection describes 1) the datasets used; 2) the registration methods to be evaluated; and 3) the evaluation criteria.

4.1.2.1. Datasets for Evaluation

Evaluations have been performed on two public datasets of normal brain MR images. All images have been skull-stripped by neuroradiologists in a pre-processing step. Skull-stripping
Table 3: The two public brain MRI datasets used in the validation – NIREP (Christensen et al., 2006) and LONI-LPBA40 (Shattuck et al., 2008) datasets. Included in both datasets are healthy subjects. They have different image sizes because the images in the NIREP dataset contain larger space for the background.

<table>
<thead>
<tr>
<th></th>
<th>NIREP</th>
<th>LONI-LPBA40</th>
</tr>
</thead>
<tbody>
<tr>
<td># subjects</td>
<td>16 (all are used)</td>
<td>40 (only 15 used)</td>
</tr>
<tr>
<td>image modality</td>
<td>T1-weighted MRI</td>
<td>T1-weighted MRI</td>
</tr>
<tr>
<td>image size (voxels)</td>
<td>256 \times 300 \times 256</td>
<td>181 \times 217 \times 181</td>
</tr>
<tr>
<td>voxel size (mm$^3$)</td>
<td>1.0 \times 1.0 \times 1.0</td>
<td>1.0 \times 1.0 \times 1.0</td>
</tr>
<tr>
<td># expert-defined ROIs</td>
<td>32</td>
<td>56</td>
</tr>
</tbody>
</table>

was done at institutions that released the datasets. Table 3 presents information concerning the two datasets. Both datasets contain T1-weighted (T1w) MR images. Moreover, they have expert-defined regions-of-interest (ROIs) that, as will be explained later in subsection 4.1.2.3 serve as references for evaluating the registration accuracy. Specifically, the NIREP dataset has expert annotations of 32 localized ROIs which are distributed in the frontal, parietal, temporal and occipital lobes, cingulate gyrus, and insula. The LONI-LPBA40 dataset has expert annotations of 56 localized ROIs in all four lobes, insula, cingulate gyrus, cerebellum, and brainstem. Detailed lists of these ROIs and how they are labeled by experts can be found in (Christensen et al., 2006) and (Shattuck et al., 2008).

Fig. 27 shows a typical set of subjects from the NIREP and LONI-LPBA40 datasets. Both their intensity images and their expert annotation images are shown. Both datasets show a considerable inter-subject anatomical variation, which is one of the most common challenges for cross-subject brain image registration.

Common pre-processing steps include N3 inhomogeneity correction (Sled et al., 1998), intensity normalization and matching, and flirt-based affine registration (Jenkinson and Smith, 2001).
Figure 27: Four typical subjects in the NIREP dataset (two rows in the top) and four typical subjects in the LONI-LPBA40 dataset (two rows in the bottom). For each subject, both the intensity image and the expert ROI annotation image are shown.
4.1.2.2. Registration Methods to be Evaluated

Twelve publicly-available methods are included in this study. They are summarized in Table 4 and reviewed in more detail in the literature review chapter (Chapter 2). We note that they are only a small fraction of the vast number of registration methods developed in the community. The pool can be expanded in the future to include other methods (such as elastix [Klein et al., 2010] and NiftyReg [Modat et al., 2010]). In general, we choose them because of the wide variety they represent. That is, they have different similarity measures, different deformation models and different optimization strategies, which are the most important components for registration algorithms (see Table 4). Out of those 12 registration methods, 9 methods were included in a recent brain registration evaluation study [Klein et al., 2009]: flirt [Jenkinson and Smith, 2001], fnirt [Andersson et al., 2008], AIR [Woods et al., 1992, 1998], ART [Ardekani et al., 2005], ANTs [Avants et al., 2008], CC-FFD [Rueckert et al., 1999], SSD-FFD [Rueckert et al., 1999], MI-FFD [Rueckert et al., 1999], and Diffeomorphic Demons [Vercauteren et al., 2009]. In addition, we have included three recently-developed registration methods that are not included in the study [Klein et al., 2009]. They are: Demons [Thirion, 1998; Vercauteren et al., 2009] (a ITK-based public software), DRAMMS [Ou et al., 2011] (our method that matches images by voxel-wise texture attributes instead of intensities and modulates voxels by mutual-saliency weightings), and DROP [Glocker et al., 2008] (a novel discrete optimization strategy that balances between the speed and the accuracy of registration).

To obtain the best deformation accuracy for each method, we adopt parameters optimized on the same datasets as collected in [Klein et al., 2009]. These parameters were optimized by the developers of the software packages evaluated in that study. Therefore, the evaluation remains fair. Note that this rule applies only to the 9 registration methods which were included in [Klein et al., 2009]. For the other 3 methods that were not included in that study [Klein et al., 2009] — Demons, DROP, DRAMMS, the parameters are set either by recommendations from the software manuals, or more often based on the feedback from
Table 4: Registration methods to be evaluated in cross-subject registration of brain images. This table is only a brief summary of them. More detail about the registration methods listed below can be found in the literature review chapter of this dissertation (Chapter 2). Abbreviations: diff.–diffeomorphism; MI – mutual information; NMI – normalized MI; SSD – sum of squared difference; MSD – mean squared difference; CC – correlation coefficient; NCC – normalized CC.

<table>
<thead>
<tr>
<th>Method</th>
<th>Deformation Model</th>
<th>Similarity</th>
<th>Regularization</th>
</tr>
</thead>
<tbody>
<tr>
<td>flirt</td>
<td>affine</td>
<td>SSD/CC/MI/NMI(used)</td>
<td>–</td>
</tr>
<tr>
<td>fnirt</td>
<td>Cubic B-spline</td>
<td>SSD</td>
<td>bending energy</td>
</tr>
<tr>
<td>AIR</td>
<td>5th polynomial warps</td>
<td>MSD</td>
<td>by polynomial</td>
</tr>
<tr>
<td>ANT's</td>
<td>symmetric diff.</td>
<td>CC</td>
<td>Gaussian smoothing</td>
</tr>
<tr>
<td>ART</td>
<td>homeomorphism</td>
<td>NCC</td>
<td>Gaussian smoothing</td>
</tr>
<tr>
<td>CC-FFD</td>
<td>Cubic B-spline</td>
<td>CC</td>
<td>bending energy</td>
</tr>
<tr>
<td>MI-FFD</td>
<td>Cubic B-spline</td>
<td>MI</td>
<td>bending energy</td>
</tr>
<tr>
<td>SSD-FFD</td>
<td>Cubic B-spline</td>
<td>SSD</td>
<td>bending energy</td>
</tr>
<tr>
<td>DROP</td>
<td>Cubic B-spline</td>
<td>MSD</td>
<td>bending energy</td>
</tr>
<tr>
<td>Demons</td>
<td>optical flow</td>
<td>SSD</td>
<td>Gaussian smoothing</td>
</tr>
<tr>
<td>Diff. Demons</td>
<td>diff. optical flow</td>
<td>SSD</td>
<td>Gaussian smoothing</td>
</tr>
<tr>
<td>DRAMMS</td>
<td>Cubic B-spline</td>
<td>SSD of attributes</td>
<td>bending energy</td>
</tr>
</tbody>
</table>

the accuracy and aggressiveness measures described in Section 4.1.2.3. After the optimal parameters for each registration method have been determined, they are kept static for all experiments in this section. Moreover, two sets of parameters, one for more aggressive and one for more smooth deformations, are used for the top-ranking methods, to offer more informed choice under diverse needs.

To avoid bias in the selection of target images in registration, we have considered all possible images as the source and the target in registration within each dataset. This results in, for each registration method, 240 (= 16 × 15) pair-wise registrations in the NIREP dataset and 210 (= 15 × 14) pair-wise registrations in the LONI-LPBA40 dataset.

4.1.2.3. Evaluation Criteria

We measure both deformation accuracy and deformation aggressiveness for each registration method. Specifically, accuracy is accessed via the Jaccard Overlap (Jaccard, 1901) between the deformed expert-annotation of the source image and the expert-annotation of the target
image in all available ROIs. Larger overlap often indicates greater spatial alignment between subjects (Klein et al., 2009; Christensen et al., 2006). Mathematically, given two regions $S$ and $T$ in a 3D space, and define the volume of a region as $V(\cdot)$, the Jaccard overlap $J(S,T)$ (Jaccard, 1901) between these two regions is defined as

$$J(S,T) = \frac{V(S \cap T)}{V(S \cup T)}.$$  \hspace{1cm} (4.1)

Some other studies used the Dice overlap coefficient (Dice, 1945) between two regions as the metric for evaluating registration accuracy, which is defined as

$$D(S,T) = \frac{2V(S \cap T)}{V(S) + V(T)}. \hspace{1cm} (4.2)$$

Actually, the Jaccard and the Dice overlap metrics show the same trend, because they are connected by

$$D = \frac{2J}{1 + J}. \hspace{1cm} (4.3)$$

Therefore, the use of any one of them is indicative of the other. In this section, we only report registration accuracies measured by the Jaccard overlap.

A deformation is considered more ”aggressive” if it leads to greater expansions/shrinkages to capture the anatomical variability between different subjects. This often leads to intersections in the obtained deformations (i.e., non-diffeomorphic). We use Jacobian determinants to measure the aggressiveness of deformations. The Jacobian determinant value measures the volumetric change ratio at a voxel. Given a deformation $h$ in a 3D space, we assume that the axes of this 3D space are indexed by $i, j, k$ in three orientations, then the Jacobian matrix at a voxel $x$ is defined as

$$\text{Jac}(x) = \begin{pmatrix}
\frac{\partial h^2(x)}{(\partial i)^2} & \frac{\partial h^2(x)}{\partial i \partial j} & \frac{\partial h^2(x)}{\partial i \partial k} \\
\frac{\partial h^2(x)}{\partial i \partial j} & \frac{\partial h^2(x)}{(\partial j)^2} & \frac{\partial h^2(x)}{\partial j \partial k} \\
\frac{\partial h^2(x)}{\partial i \partial k} & \frac{\partial h^2(x)}{\partial j \partial k} & \frac{\partial h^2(x)}{(\partial k)^2}
\end{pmatrix}. \hspace{1cm} (4.4)$$
and the determinant of this Jacobian matrix (often known as "Jacobian determinant") is

\[ \text{JacDet}(x) = \det(\text{Jac}(x)). \] (4.5)

The Jacobian determinant is greater than 1 for volume expansion, between 0 and 1 for volume shrinkage, 0 for volume vanish and smaller than 0 if self-folding occurs. In particular, we measure 4 Jacobian-based metrics: 1) the number of deformations with negative Jacobian determinants (or equivalently, the percentage of cases which show self-intersections); 2) the percentage of voxels having negative Jacobian determinants; 3) the minimum and 4) the maximum Jacobian determinants in a deformation. Finally, we use one metric, the range of Jacobian determinants (=\(\text{maxJac}-\text{minJac}\)), to quantify deformation aggressiveness.

Another metric we can use is Harmonic energy, or bending energy of the deformation field (Bookstein 1989). The trend is similar in our experiments.

Jacobian calculation is not a trivial task. Theoretically, Jacobian is derived from continuous space. However, because of the discrete nature in the digitalized medical images, all calculations of Jacobians are only numerical approximations. Another difficulty is that different software packages calculate Jacobians in different numerical fashions. For fairness, we use the Jacobian calculator from the ITK library (Ibanez et al. 2003), which is arguably the most commonly-used Jacobian calculator in the field. This requires converting deformation files from different software packages into a standard ITK-compatible MetaImage format. We performed careful checks to ensure that the conversion process reproduces identical warped images.

### 4.1.3. Results and Observations

A typical set of DRAMMS registration results are visualized in Fig. 28. The results are from the registration between two NIREP subjects. Visually, DRAMMS has led to plausible
alignment from the source to the target images, with the registered image now in anatomical correspondence with the target image. The same deformation obtained from registering the source to the target intensity images is used to warp the expert annotation image from the source to the target space. Based on such a warping, we can quantify the registration accuracy.

Quantitative validation results (accuracy, aggressiveness, and their correlation) are presented below for both the NIREP and LONI-LPBA40 datasets. Observations follow each set of results.

A. Deformation Accuracy

Fig. 29 shows the ROI-overlap-indicated deformation accuracy in the NIREP and LONI-LPBA40 datasets. The accuracy is averaged among all ROIs available in each dataset. In this figure, a parenthesis (A) denotes the aggressive version of this algorithm, i.e. a version with very low regularization/smoothness. Several observations can be made from Fig. 29:

A1) DRAMMS(A) (the aggressive version of DRAMMS), and ANT(A), (the aggressive version of ANTs), score the highest Jaccard overlap in both datasets. Following them are DROP(A), Demons and ART registration methods. Of these methods, ANTs and ART were included in (Klein et al., 2009) evaluation study and were found to be two most accurate methods. Our findings here show the same trend. In addition, the three recently-developed methods — DRAMMS, DROP and Demons, which were not included in (Klein et al., 2009) study, show competitive performances.

A2) Methods such as SSD-FFD, fnirt, DROP use intensity differences (SSD) as the similarity metric. On average, they have reasonable Jaccard overlaps. However, they also show larger variations, and therefore are less stable than methods using CC, MI or attribute-based similarity measures. This outlines the advantage of using attributes other than intensity
Figure 28: Registration of 3D brain MR images from two different NIREP subjects by DRAMMS.
differences to more consistently capture cross-subject anatomical variations.

A3) The sophistication of the deformation mechanism (such as the FFD model as used in DRAMMS and the diffeomorphism model as used in ANTs) is another factor contributing to the higher registration accuracy, compared to relatively simpler deformation mechanisms (such as 5-th order polynomials in AIR).

B. Deformation Aggressiveness

The aggressiveness of a deformation as indicated by Jacobian statistics is presented in Figs. 30 and 31 for the NIREP and LONI-LPBA40 datasets, respectively. In each figure, the sub-figures from left to right, and from top to bottom, are: the number of deformations with negative Jacobian determinants; the percentage of voxels with negative Jacobian determinants; the minimum; and the maximum Jacobian determinants in deformations. We observe the following from these results:

B1) From the top row in both figures, fnirt is the only non-rigid registration method that guarantees diffeomorphism in this dataset. Diffeomorphism implies the absence of any negative Jacobian determinants (i.e. no self-folding) in deformations. It preserves topology and one-to-one forward and backward correspondences. The fnirt method guarantees diffeomorphism by directly checking and removing negativities in the Jacobian map. However, this is at the cost of overlap-indicated registration accuracy, as reflected in Fig. 29. Though attractive from a mathematical perspective, in general, it is unclear whether a diffeomorphism is a desirable property in an anatomical sense, especially in the presence of large inter-subject anatomical variations. In particular, diffeomorphism implies that every single piece of tissue in one brain has a unique counterpart in any other brain, which is certainly an assumption that would be difficult to defend with theoretical or experimental arguments.

B2) DRAMMS(A), ANTs(A), DROP(A), Demons and ART(A) score higher overlaps, indicting higher registration accuracies (Fig. 29). Interestingly, the results in the lower row
Figure 29: Box-and-Whisker plots of deformation accuracies in the NIREP and LONI-LPBA40 datasets. Letter "A" stands for aggressive version and "S" for smooth version of a method.
of the aggressiveness figures (Figs. 30 and 31) show that these methods have quite different deformation styles. In particular, DRAMMS(A) and ART(A) have greater maximum Jacobian determinants, trying to capture the individual variability with larger expansions. ANTs(A) and ART(A) have more negative minimum Jacobian determinants, trying to capture the same individual variability with slightly more self-foldings in deformations. In the absence of ground-truth on point-to-point correspondence, it is hard to draw any conclusions on the two different deformation styles.

C. Correlation between Accuracy and Aggressiveness Surrogates

The surrogates for the accuracy of deformations (i.e., Jaccard overlap) and the surrogates for the aggressiveness of deformations (i.e., Jacobian range) are plotted jointly in Fig. 32. Here, the y-axis is the mean Jaccard overlap over all 32 (56) structures and all 240 (210) registrations in the NIREP (LONI-LPBA40) dataset, indicating the overall accu-
Figure 31: Deformation aggressiveness as indicated by Jacobian-based metrics in the LONI-LPBA40 dataset. Upper left: number of deformations (out of all 210) that have negative Jacobian determinants; Upper right: box-and-whisker plot of percentage of voxels having negative Jacobian determinants in a deformation; Lower row: min (left) and max (right) Jacobian determinants.
racy of a registration method. The x-axis is the mean range of Jacobian determinants \(=\text{mean}(\max\text{JacobianDet}-\min\text{JacobianDet})\) over all 240 (210) registrations, indicating the aggressiveness of a method. Three observations can be made from this figure:

**C1)** Generally, methods score higher overlap when allowing more aggressive deformations.

**C2)** An ideal registration method should obtain the highest possible overlap while being diffeomorphic to preserve the topology in images. Combining Fig. 29 for the accuracy of deformations with the upper left parts of Figs. 30 and 31 for the aggressiveness of deformations, DRAMMS(A) and ANTs(A) have achieved the highest accuracy while both having a reasonably low percentage of voxels with negative Jacobian determinants (< 0.2%).

**C3)** Overall, in these datasets, DRAMMS(A) shows an equivalent level of accuracy to ANTs(A), which was found to be the most accurate method in (Klein et al., 2009), in cross-subject brain MR image registration. Their aggressiveness levels are also comparable. Moreover, their smooth versions, DRAMMS(S) and ANTs(S), while much smoother and carrying far fewer or almost none negative Jacobian determinants, have also achieved high accuracies that are in general comparable to other high-ranking methods such as DROP, ART, and Diffeomorphic Demons.

**4.1.4. Discussion**

This section validates DRAMMS among 12 voxel-wise registration methods within the context of cross-subject brain MR image registrations in two public datasets. Results show that DRAMMS has obtained almost equivalent deformation accuracy with a similar level of deformation aggressiveness compared with ANTs, which was found to be the most accurate method in a recent third-party evaluation study ((Klein et al., 2009)) that involved 14 registration methods. Also, DRAMMS and ANTs are consistently more accurate than the other registration methods evaluated here in this section, followed by DROP, Demons and
Figure 32: A joint plot of deformation accuracy and aggressiveness surrogates in the NIREP and LONI-LPBA40 datasets. Letter "A" stands for aggressive version and "S" for smooth version of a method.
ART methods.

The validation framework in this section has three features: 1) more recent registration methods (DRAMMS, Demons, DROP) are included in the evaluation, and are shown to be highly accurate; 2) deformation aggressiveness is measured in addition to deformation accuracy, revealing their strong correlation; 3) two sets of parameters, one set corresponding to more aggressive deformations and one set corresponding to smoother deformations, are evaluated for those most accurate methods, establishing their performances at a range of deformation aggressiveness levels.

This section has focused on skull-stripped brain images. Skull-stripping is a common pre-processing step in brain image analysis. It helps focus on cortical and sub-cortical regions that are of most interest in many longitudinal studies (e.g. brain aging) and population studies (e.g. abnormality diagnosis from the normal cohort). Therefore, registration of skull-stripped brain images is perhaps the most extensively studied field in the image registration community. The validation in this section has established the accuracy of DRAMMS in this popular area.

In the following sections, DRAMMS will be further evaluated in more challenging brain registration tasks, such as registration in the presence of skull and background noise, in the presence of lesions, and even in the presence of recurred tumors. These tests will further establish the accuracy and the robustness of DRAMMS in different brain registration tasks.
4.2. In Registering Raw Brain MR Images Across Subjects

4.2.1. Introduction

This section evaluates DRAMMS in cross-subject registration of raw brain images. Raw brain images are those obtained from imaging devices which have not yet gone through any pre-processing steps. Registration of raw brain images is typically required in two situations: 1) when skull-stripping, a major pre-processing step, is unavailable (possibly because manual skull-stripping is too tedious, or because automatic skull-stripping fails, which is not unusual); 2) as part of the skull-stripping process itself — for example, in a multi-atlas based skull-stripping framework (such as our motivational study 4 in Section 1.1), when raw images of one or more subjects (serving as atlases) need to be registered onto the raw image of the target subject to infer the brain mask in the target image.

Registering raw brain images usually involves one or more of the following challenges. In order of increasing difficulty, some typical challenges include:

1. the background noise and magnetic field inhomogeneities, which may still remain even after noise reduction and inhomogeneity correction;

2. the anatomical variations, including not only the variability in brain structures, but also in non-brain (often off-focal) structures, such as the skull, dura, nose, ear, neck, and others;

3. the imaging angle differences (due to the difficulty in automatic AC-PC alignment); and worse

4. imaging scope differences, or the field-of-view (FOV) differences in images of different
subjects. An example of the FOV difference can be seen in Fig. 34 where the nose, face and neck are included in the image frame (FOV) of an ADNI subject (the third row, the subject on the left) but are not included in the image frame (FOV) of an OASIS subject (the second row, the subject on the left). In such situations, the FOV difference causes missing correspondences between the raw images of two subjects. This happens more often between multi-institutional datasets (cross-dataset), but may also happen within a mono-institutional dataset (within-dataset). For instance, see the FOV difference between the two subjects from the same ADNI dataset (the third row of Fig. 34).

Because of the above challenges, cross-subject registration of raw brain images is generally considered a fairly challenging problem. For this reason, a skull-stripping step often precedes brain registration. However, registration of raw brain images can be part of the (atlas-based) skull-stripping process itself. In this circumstance, registration of raw brain images cannot be bypassed.

On the other hand, the DRAMMS framework is believed to be robust to many of the challenges above, when compared with the use of many existing intensity-based methods in registering raw brain images. One reason is that the attribute matching component of DRAMMS should make registration less sensitive to the influence of noise and inhomogeneities. The attribute matching component should also be more accurate in capturing large anatomical variations, as has already been shown in the previous section concerning registration of skull-stripped brain images. More importantly, the mutual-saliency component in DRAMMS should provide more robustness with regard to the angle and FOV differences, or the missing correspondences problem, compared with other voxel-wise methods that use all imaging voxels equally.

Therefore, this section validates the performance of DRAMMS in cross-subject registration of raw brain images. Four registration methods are compared in two scenarios. Scenario 1 is to validate the alignment of the entire image (i.e., brain and non-brain tissues types or
structures). In this scenario, one public dataset is used, which has 11 expert-defined ROIs. Scenario 2 focuses on the brain only — to validate the alignment of the brain masks. In this scenario, three public datasets from different institutions are used. Registrations are evaluated in both within-dataset registration, which is relatively easier, and cross-dataset registration, which is considerably more difficult mainly due to the missing correspondences problems caused by FOV differences in multi-site data. The validation results in these two scenarios will establish DRAMMS as a favorable method for the accuracy and robustness in registering images from single- and multi-institutions.

In the following, Subsection 4.2.2 describes the validation protocols; Subsection 4.2.3 presents the validation results and observations; and finally Subsection 4.2.4 discusses and concludes this section.

4.2.2. Validation Protocol

This subsection describes 1) the datasets; 2) the registration methods to be evaluated; and 3) the evaluation criteria. The pipeline is similar to that outlined in Section 4.1.2 so this section focuses primarily on the different elements.

4.2.2.1. Datasets for Evaluation

We have used 4 public datasets for two registration scenarios involving raw brain images.

The first scenario focuses on the alignment of the entire image (including various brain and non-brain tissue types and structures). In this scenario, the public BrainWeb dataset is used (Aubert-Broche et al., 2006). The BrainWeb dataset contains raw brain images of 20 healthy subjects. In each subject, every image voxel has been annotated as one of 11 brain and non-brain tissue type and structures: CSF, Gray Matter, White Matter, Fat, Muscle, Muscle/Skin, Skull, vessels, around fat, dura matter, and bone marrow. Fig. 33
Figure 33: Three typical subjects from the BrainWeb dataset. Every subject has a raw intensity image and the corresponding annotation image with 11 brain or non-brain structures. This dataset is used to validate how the entire image (brain and non-brain structures) has been aligned.

presents 3 typical subjects from the BrainWeb dataset, including their intensity images and corresponding annotation images. We have randomly picked 11 such subjects for validation, leading to 110 \((11 \times 10)\) pair-wise registrations for each registration method under comparison.

For the second scenario, which focuses purely on the brain and examines the alignment of brain masks, three public datasets have been used. They are popular datasets for validating skull-stripping methods (e.g., Iglesias et al. (2011)). They also present different levels of difficulty, as we will see in the results section. The first is the Internet Brain Segmentation Repository (IBSR) dataset (Tsang et al., 2008). The IBSR dataset consists of 20 T1-weighted scans from the Center for Morphometric Analysis at the Massachusetts General Hospital. In this dataset, the brain was manually delineated by trained investigators in all scans. The second is the OASIS dataset (Marcus et al., 2007), which consists of 55 females and 22 males, with \(51.64 \pm 24.67\) years of age. The brain masks were first generated by an automated method based on registration to an atlas, and then proofread and corrected by human experts before the release. The third dataset is from the baseline images of normal control subjects in the Alzheimer’s Disease Neuroimaging Initiative (ADNI) study (http:}
For a more computationally feasible validation, five subjects are randomly selected from each of these three datasets. This leads to a total of 60 \(((5 \times 4)/\text{dataset} \times 3\text{datasets})\) within-dataset registrations and a total of 150 \((15 \times 10)\) cross-dataset registrations for each method under evaluation.

Fig. 34 presents some representative subjects (images and brain mask annotations) from the three datasets used in this section (IBSR, OASIS, ADNI). Experiments are performed both within-dataset and cross-dataset. The registration process faces severe challenges posed by the large anatomical variation, background noise, contrast differences, and large FOV differences and hence the missing correspondences problems both within- and especially across- dataset. Visually it is not difficult to see that the IBSR and ADNI datasets are more difficult for within-dataset registration because of their within-dataset FOV differences. Such differences are not present in the OASIS dataset, and therefore, within-dataset is expected to be less difficult in the OASIS dataset. For all cross-dataset registrations, FOV difference is a major challenge.

### 4.2.2.2. Registration Methods to be Evaluated

We start with the 12 registration methods that are introduced in Section 4.1.2.2. All 12 methods are tested in registration scenario 1, which examines how well the entire image, including the brain and non-brain structures, has been aligned. Then, in scenario 2, which focuses on the brain only, we validate DRAMMS against 3 other representative methods (ANTs, Demons and fnirt).

For fairness, the same parameter settings used for registration of skull-stripped brain images in Section 4.1 are adopted here. In other words, the parameters for each registration method remain unchanged between registering images with the skull \(i.e.,\) raw images) and registering images without the skull.
Figure 34: Typical raw brain images from the IBSR, OASIS and ADNI datasets. Presented here are two representative subjects from each dataset. Every subject has both raw intensity image and the corresponding brain mask next to it. Image registration is challenged by the large anatomical variation, background noise, contrast difference, and worse, large FOV differences within and especially across datasets.
4.2.2.3. Accuracy Criteria

In registration scenario 1, which looks at various brain and non-brain structures in the entire image, we measure deformation accuracy by the Jaccard overlap between the warped ROI annotation and the target ROI annotation. A higher Jaccard overlap usually indicates a higher registration accuracy.

In registration scenario 2, which focuses on brain masks only, two criteria are used to evaluate deformation accuracy: 1) the Jaccard overlap between the warped brain mask and the target brain mask — a higher overlap indicating a higher registration accuracy; 2) the 95-percentile distance error between the warped and the target brain masks — a smaller distance error indicating a higher registration accuracy. Fig. 35 presents the definition of the distance error. It is a measure of the spatial distance between the boundary of the warped ROI and the boundary of the target ROI at each boundary point. The additional use of distance error is because that, unlike other localized small-scale regions, the brain mask is a big ROI; and the Jaccard overlap on big ROIs may not necessarily indicate the discrepancies in the boundary of the ROIs. For example, two brain masks may have a large disagreement in some boundary areas although they may have a high Jaccard overlap, simply because the brain mask is so big. Moreover, to avoid any outliers, we only measure 95-percentile of distance error, instead of the max distance error.

4.2.3. Results and Observations

First, the registration accuracy is presented for the entire raw images (brain and non-brain structures) in Scenario 1. After that, Scenario 2 focuses on the brain only, for both within-dataset and cross-dataset registrations.

Scenario 1. Registration Accuracy in the Entire Raw Image

Fig. 36 shows the ROI-overlap-indicated registration accuracy in the BrainWeb dataset.
The 11 ROIs in the BrainWeb dataset include the brain, the skull as well as other non-brain tissue types and structures (as detailed in 4.2.2.1). Therefore, the average accuracy among all ROIs is indicative of registration accuracy in the entire raw image. Several observations can be made:

a) DRAMMS(A), the aggressive version of DRAMMS, DROP(A), and Demons demonstrate similarly high accuracy in the registration of the entire images, followed closely by the ANTs(A) and ART(A) methods. These registration methods are also among the most accurate ones in registering skull-stripped brain images.

b) On the other hand, the average Jaccard overlap by DRAMMS(A) is only around 0.32 in registering raw brain images (Fig. 36) compared to the average Jaccard of 0.52–0.57 in registering skull-stripped brain images (Fig. 29). This clearly underlines the increased difficulty in registering raw brain images.

**Scenario 2. Registration Accuracy in Brain Masks — Within and Across Datasets**
In this registration scenario, we focus on the accuracy of registration in warping the brain masks. This is the basis for the registration-based skull-stripping framework. Results are shown for both within-dataset registration and cross-dataset registration. For within-dataset registration, Fig. 37 presents some visual results from DRAMMS. Fig. 38 shows the accuracies of 4 competing registration methods (ANTS, fniRT, Demons, and DRAMMS). For cross-dataset registration, Fig. 39 presents some visual results from DRAMMS. Fig. 40 shows the accuracy of the same four competing methods. Several observations can be made:

a) The Jaccard regional overlap and the distance error show the same trend for indicating registration accuracy;

b) In within-dataset registrations, ANTs scores the highest accuracy in the OASIS dataset. The OASIS dataset presents the lowest level of inter-subject FOV differences compared with other datasets used in this section (Fig. 34). In the other two datasets (IBSR, ADNI), where subjects within the dataset do exhibit large FOV differences and hence larger regions
of missing correspondences, DRAMMS scores the highest accuracy. Demons is close to ANTs and DRAMMS in all datasets. Overall DRAMMS has the highest median accuracy and the smallest spread of accuracy, demonstrating its robustness towards the anatomical variability in different datasets.

c) Cross-dataset registration of raw brain images is a considerably more difficult problem than within-dataset registration. This is because of the large anatomical, intensity, and contrast differences, and especially the large FOV differences in subjects from different datasets (see Fig. [34]). Registration results clearly reflect such difficulties. In the cross-dataset registration results presented in Fig. [40], registration accuracies decrease for every method, compared with the accuracies of within-dataset registrations (Fig. [38]). However, the accuracy decreases to different extents. fnirt is most affected, with a big decrease from the median Jaccard of close to 0.7 in within-dataset registrations to the median Jaccard of 0.2, or mostly failures, in cross-dataset registrations. ANTs and Demons are less influenced, but still lose about 0.1 in the median Jaccard overlap moving from within-dataset to cross-dataset registrations. DRAMMS is least influenced, losing about 0.06 in the median Jaccard overlap. Overall, DRAMMS retains the highest accuracy in cross-dataset registrations of raw brain images, showing the highest level of robustness with regard to large anatomical
Figure 38: Registration accuracy on brain masks for all 60 within-dataset registrations in the IBSR, OASIS and ADNI datasets. Top row: the accuracy indicated by the Jaccard overlap between the warped and the target brain mask. Bottom row: the accuracy indicated by the 95-percentile distance error between the warped and the target brain masks.
Figure 39: Typical DRAMMS registration results in cross-dataset registrations. The source image is from the OASIS dataset and the target image is from the ADNI dataset. DRAMMS is able to handle the inter-subject differences in the intensity distribution and anatomy, and especially the large difference in FOV, which is a common yet fairly challenging problem in registration of raw brain images.

Figure 40: Registration accuracy on brain masks for all 150 cross-dataset registrations. Top row: accuracy indicated by the Jaccard overlap between the warped and the target brain mask. Bottom row: accuracy indicated by the 95-percentile distance error between the warped and the target brain masks.
variations, and especially FOV differences in multi-institutional datasets.

**d)** Overall, we can expect 0.93 and 0.87 median Jaccard overlap in within- and cross-dataset registration of raw brain images by DRAMMS in the datasets tested in this section. This provides a solid foundation for (multi-)atlas-based skull-stripping as will be further presented in Chapter 5.

### 4.2.4. Conclusion

This section validates DRAMMS in registering raw brain images, both within- and cross-datasets. Experiments are conducted in four public datasets with expert-defined ROIs.

Overall, DRAMMS, ANTs, Demons and DROP show the highest accuracy compared with the other registration methods in this section. Specifically, DRAMMS, DROP and Demons are more accurate in aligning brain and non-brain structures and tissue types in the BrainWeb dataset. In within-dataset registrations, ANTs demonstrate an higher accuracy for brain masks in 1 dataset (OASIS) and DRAMMS in the remaining 2 relatively more difficult datasets (IBSR and ADNI). In cross-dataset registrations, DRAMMS scores consistently higher accuracies. The trend is that, as registration difficulty increases, accuracy decreases for all registration methods, but DRAMMS is least affected compared with the other three methods.

Aligning brain masks in raw brain image registration lays the foundations for (multi-)atlas-based skull-stripping. In Chapter 5, we will apply the herein validated DRAMMS to skull-stripping of raw brain images.
4.3. In Registering Lesion-bearing Brain to Healthy Brain MR Images

4.3.1. Introduction

This section validates DRAMMS in registering lesion-bearing brain images to a healthy brain template image. In population studies such as our motivational study 1 (Section 1.1), lesion-bearing images from a cohort of patients need to be spatially normalized, or registered, into the same anatomical space before they can be jointly analyzed for function-structure correlation, for population statistics, and for tracking treatment effects. Here, image registration plays a central role. The main challenge, however, is the missing correspondence problem — lesions appearing in the patient image but not in the healthy template image.

A popular approach to deal with the lesion-induced missing correspondence problem is the so-called cost-function-masking (CFM) strategy [Brett et al., 2001]. CFM masks out lesion regions from image registration cost functions. As a result, correspondences will only be established in other normal regions, and the lesion regions will move by a smooth interpolation from the movements of the neighboring normal regions. The immediate advantage is the removal of the negative impact of the lesion regions, where correspondences could be hardly established. However, the limitation is the need for an accurate lesion delineation, which is usually conducted by human experts. Because of this, the CFM strategy is usually only semi-automatic.

On the other hand, DRAMMS may suit the task of registering lesion-bearing brain images to normal brain images. The mutual-saliency component in DRAMMS provides automated differentiation between regions with correspondences and regions without correspondences in the other images. Without requiring pre-segmentation of lesions, it helps reduce the
impact of lesion regions, and drive the registration mainly by other normal regions.

Therefore, this section validates DRAMMS in registering lesion-bearing brain images to normal brain images. DRAMMS is compared with other general-purpose registration methods, and in particular with other methods equipped with the CFM strategy that is specifically designed for registration of brain images with pathologies.

The remainder of this section is organized as follows. Subsection 4.3.2 describes the validation protocol (data, methods, criteria); Subsection 4.3.3 presents the validation results; and finally Subsection 4.3.4 discusses and concludes this section.

4.3.2. Validation Protocol

This subsection describes 1) the datasets; 2) the registration methods to be evaluated; and 3) the evaluation criteria.

4.3.2.1. Datasets for Evaluation

There are almost no public datasets available for evaluating lesion brain registrations. Therefore, simulation is employed in this section. Real data are not used for two reasons: 1) obtaining the ground truth deformations is difficult on real data; 2) real data containing brain tumors will be presented in the next section (4.4), with similar registration settings and even an higher level of registration difficulty.

Fig. 41 presents the simulated brain images with lesions and the registration setup. First, six real T1-weighted brain MR images are selected from the Baltimore Longitudinal Study of Aging (BLSA) dataset (image size $256 \times 256 \times 124$ and voxel size $0.9375 \times 0.9375 \times 1.5mm^3$). They are all from healthy subjects. One is randomly selected as the common "template" image that the other five images will be registered into. Then, in these other five images, we simulated lesions of five levels of increasing sizes (at the rate of 5mm increment in 3D
Figure 41: Brain images with simulated lesions. Red arrows point to the simulated lesions of increasing sizes for each subject image.

radius). The simulated is via manually "painting" the voxels in the normal regions with lesion-like intensities. In this setup, we can perform subject-to-template registration with and without simulated lesions. The results will help evaluate the response of a registration method to lesion-induced missing correspondences.

4.3.2.2. Registration Methods to be Evaluated

Eight registration methods are included in this evaluation. They include six general-purpose registration methods (DRAMMS, ANTs, ART, Demons, DROP, and fnirt) and two task-specific methods (ANTs+CFM and fnirt+CFM). Of those eight methods, DRAMMS, ANTs, ART, Demons and DROP are found to be most accurate in registering normal brain images, and ANTs+CFM and fnirt+CFM are the ANTs and fnirt methods equipped with
the CFM strategy.

For fairness, the optimal parameters in normal brain registrations are adopted here for lesion brain registration for every registration method. This will reveal the difference of the a registration method’s performance for registering images with and without lesions.

The CFM strategy requires an input of lesion masks. Usually lesion masks are delineated by human experts. In our experiment with simulated data, we input the true lesion masks as they are already available during the simulation. We therefore evaluate the CFM strategy in the absolutely most favorable situation.

4.3.2.3. Evaluation Criteria

Unlike brain tumors, vascular lesions replace the normal anatomy, but do not have mass effect, and do not push or infiltrate neighboring normal regions. Therefore, an ideal registration method that is robust to the presence of brain lesions should obtain identical deformations when registering images with or without lesions. In practice, the obtained deformation does change when lesions are present. Therefore, we can measure the robustness of a registration method by the magnitude of deformation difference (MDD) between the deformations obtained in the registration with or with the presence of lesions. A small MDD implies a high level of robustness towards the presence of lesions. Moreover, the change rate of MDD when the size of lesions increases reflects the stability of a registration method.

Mathematically, we denote a registration method as $M$, the common normal template image as $T$, and a subject image as $S_{i,l}$. Here $i$ indexes the five different subject images we randomly selected from the BLSA dataset. $l$ indexes the lesion size in subject image $S_i$. $l = 0$ means there is no lesion in the subject image, such as the first column in the image blocks in Fig. [11]. In this case, registration between $S_{i,l=0}$ and the normal template $T$ is simply an cross-subject registration of normal skull-stripped images, whose accuracy has
already been measured in Section 4.1. When \( l = 1, 2, \ldots, 5 \), this means subject image \( S_{i,l} \) has lesions of increasing sizes. In such cases, the registration between \( S_{i,l}(l = 1, 2, \ldots, 5) \) to \( T \) is a registration between a lesion-bearing brain image to the normal template. The magnitude of the deformation difference, \( \text{MDD}(M, l) \), for a registration method \( M \) under lesion size index \( l \) is defined as

\[
\text{MDD}(M, l) = \int_{x \in T} |f_{i,l \neq 0}^M - f_{i,l=0}^M|dx
\]  

(4.6)

where \( f_{i,l \neq 0}^M : S_{i,l \neq 0} \rightarrow T \) is the deformation calculated from images with lesions of size \( l \), and \( f_{i,l=0}^M : S_{i,l=0} \rightarrow T \) is the deformation calculated from images without lesions. We further normalize the MDD values by the number of voxels in the template image space. The normalized MDD is reported in the results section. It quantifies the lesion-induced average displacement error at each voxel.

### 4.3.3. Results and Observations

Fig. 42 shows the normalized MDD with regard to the size of the simulated lesions, for different registration methods. Several observations can be made:

a) As expected, deformation errors increase as the size of lesions increases.

b) The cost-function-masking (CFM) strategy does reduce the negative impact of lesion-induced missing correspondences, when comparing fnirt+CFM with fnirt, and comparing ANTs+CFM with ANTs.

c) Overall, DROP and DRAMMS show a higher level of robustness in the presence of lesions.
Figure 42: Deformation errors caused by the presence of lesions. The normalized magnitude of deformation difference (normalized-MDD) measures the average voxel-wise displacement errors caused by brain lesions.

4.3.4. Conclusion

This section evaluates the performance of DRAMMS in registering brain lesion images to normal brain images. The lesion-induced missing correspondences pose a major challenge to registration. Through simulated lesions of increasing sizes, the robustness of a registration method is measured by the difference between registration with and without lesions. Results show that DROP and DRAMMS are relatively robust to lesion-induced missing correspondences. With this validation, DRAMMS can be applied to studying the patient population with brain lesions, as mentioned in the first chapter of this dissertation.

Only simulated data are used in this section. In the next section, DRAMMS will be evaluated using real data containing recurred brain tumors. A similar issue there is the tumor-induced missing correspondences. However, it will be more challenging because of the infiltrative nature of the brain tumors.
4.4. In Registering Tumor-Recurring Brain to Healthy Brain MR Images

4.4.1. Introduction

This section validates DRAMMS in registering brain images containing recurred tumors to a normal brain template image. Such a registration is motivated by Study 2 in Section 1.1. In this study, a population of patients who have recurring tumors after surgery are collected. Their brain images need to be normalized, or registered, into the same image space before they can be jointly analyzed. Image registration plays a central role, but encounters challenges mainly by tumor-induced missing correspondences.

To remedy the tumor-induced missing correspondence problem, there are three major categories of registration methods developed in the literature. One category is to convert tumor-to-normal brain image registration into tumor-to-tumor image registration. Those approaches usually seed, or simulate, tumors in normal brain images. The simulated tumors are geometrically similar to the actual tumor in the patient image (Dawant et al., 1999; Cuadra et al., 2004; Mohamed et al., 2006; Zacharaki et al., 2008; Gooya et al., 2011). The merit of these approaches is the demonstrated high accuracy and the drawback is the computation burden and the uncertainties in the tumor simulation. The second category is to convert tumor-to-normal registration into normal-to-normal registration, by replacing tumor regions in the patient image with normal-appearing intensities (Sdika and Pelletier, 2008). This approach is theoretically reasonable, but defining the exact tumor boundary for the intensity impainting remains a tough task. The third category is to mask out tumor regions from the registration cost function, and hence the name cost-function-masking (CFM) approaches (Brett et al., 2001; Andersen et al., 2010). CFM approaches are usually semi-automatic, in that the tumors to be masked out are usually manually delineated by
The above three categories of methods are mainly developed for images containing primary tumors, or the initial occurrence of tumors. Unlike the primary tumors, the recurrent tumors, which are of interest in this dissertation, are often associated with a blood pool that is accumulated after the resection of brain tumors. The blood pools cannot be modeled as part of the tumor, therefore making tumor seeding or tumor segmentation methods inappropriate. Because of this, the first two categories of methods may not directly apply to images containing recurrent tumors. CFM approaches may still apply, provided that the recurrent tumors and the blood pool are both masked out from registration cost functions. But the mask-out process often needs human expert’s delineation, which is time-consuming and irreproducible.

On the other hand, the mutual-saliency component makes DRAMMS a good candidate for fully-automatic registration of images containing recurrent tumors. The mutual-saliency values, which are calculated automatically, can help register the normal regions well and leave smooth deformation interpolations in the recurrent tumor regions. Therefore, the goal of this section is to evaluate the performance of DRAMMS in registering tumor-recurring images with a normal brain template. Ideally we would show that DRAMMS performs as well as human experts in such registration tasks. To demonstrate this, experiments are carried out on real images of 8 brain tumor patients. The results are evaluated by comparing with multi-rater annotations of ROIs and landmarks. Three other registration methods are also evaluated. In addition, registration accuracy is measured hierarchically in various image zones — within, close to, and far away from tumor recurrence regions, and the whole image. This reveals how tumor recurrence has affected registration accuracy locally and globally, from neighboring regions to far-away regions.

In the following, Subsection 4.4.2 describes the validation protocol (data, methods compared, and criteria); Subsection 4.4.3 presents validation results; and finally Subsection 4.4.4 concludes this section.
4.4.2. Validation Protocol

This subsection describes below 1) the datasets; 2) the registration methods to be evaluated; and 3) the evaluation criteria.

4.4.2.1. Datasets for Evaluation

T1-weighted MR images from 8 patients having post-surgery recurrent brain tumors are collected from the Department of Radiology at the University of Pennsylvania (image size 192×256×192, voxel size 0.977×0.977×1.0mm³). They are registered to a common, normal T1-weighted MR template image (image size 256×256×181, voxel size 1.0×1.0×1.0mm³). Fig. 3 in Chapter 1 shows some typical images in this dataset. They are all skull-stripped first by the automatic method BET (Smith, 2002) and then proofread by human raters.

For measuring registration accuracy, multi-rater annotations of tumor recurrence ROIs and landmarks throughout the image were collected. Two radiologists (Drs. Michel Bilello and Hamed Akbari) first worked together to delineate tumor recurrence ROIs in these 8 patient images. They also worked together to pick up 10 and 40 landmarks in the normal regions close to and far from tumor recurrence regions in each of these 8 patient images. We consider "close-by" normal regions as the normal regions within 30mm distance to the tumor recurrence regions, and "far-away" normal regions otherwise. After that, two raters independently mapped the same ROIs and landmarks from each patient image into the template image space. This helps observe differences between independent raters in different regions of the image space. It also provides basis for evaluating the accuracies of image registration methods.
4.4.2.2. Registration Methods to be Evaluated

Four registration methods are evaluated — DRAMMS, ANTs, Demons, and fnirt. In future studies, ANTs+CFM and fnirt+CFM approaches that are specifically designed for lesion or tumor image registrations will also be included. The CFM strategy is not compared at this point for two reasons: 1) the CFM strategy requires extra efforts for segmenting recurrent tumors and blood pools and masking them out from registration, which is not a trivial task, and is largely variable by different raters, making the comparison less objective; 2) in the current validation framework, we have already included multi-rater data; so results have already shown the performance of DRAMMS in comparison to human experts.

The optimal parameters in cross-subject normal brain image registration are adopted here. Those parameters are not necessarily optimal for the specific task of registering tumor-bearing images. However, they still provide useful information — the performance of the same set of parameters in different brain registration tasks (normal, lesion-bearing, tumor-bearing). Actually, as the results show in the later sections, these parameters have achieved accuracy quite comparable to human experts.

4.4.2.3. Evaluation Criteria

The accuracy of a registration method is measured by comparing with multi-rater annotations of ROI and landmarks. We measure registration accuracy hierarchically in different zones as sketched in Fig. 43. Those zones help establish the influence of tumor recurrence, from near to far, in local and global scales.

1. **Zone 1: tumor recurrence regions.** Registration accuracy in Zone 1 is measured by two metrics: 1) the regional overlap — the average Dice overlap of the algorithm-warped recurrent tumor ROI and two rater-warped recurrent tumor ROIs in the template space; 2) the surface distance — the average distance error (as defined in Fig. 35) between the algorithm-warped recurrent tumor ROI and two rater-warped
Figure 43: Measure registration accuracies in different zones. The solid contour filled with yellow texture denotes tumor recurrence. Zone 1-3: within, close-to, and far away from recurrent tumor region; Zone 4: brain boundary. The definition of the zones can be found in the main context in Section 4.4.2.2.
recurrent tumor ROIs in the template space. A higher regional overlap and a smaller surface distance point to a higher registration accuracy in the recurrent tumor regions.

2. **Zone 2: normal regions close to the recurrent tumor.** Normal regions close to tumor are defined as the regions within 30 mm distance to the tumor recurrence masks in the patient space. These regions, although normal-appearing, are still affected by tumor-induced missing correspondences. Registration accuracy in this zone is an important surrogate for the robustness of a registration method. Two raters, as well as algorithms, each independently find landmarks in template space that are corresponding to the common set of 10 landmarks in zone 2 in the subject space. The average Euclidean distance between algorithm-calculated corresponding landmarks and rater-labeled corresponding landmarks in the template space is used to measure registration accuracy in the normal regions close to the recurrent tumor. Smaller landmark errors point to a higher registration accuracy in the normal regions close to recurrent tumors.

3. **Zone 3: normal regions far away from the recurrent tumor.** Normal regions away from tumor are defined as the regions with more than 30 mm distance from tumor recurrence masks in the patient space. Further away regions should be less or almost not affected by tumor-induced missing correspondences if a registration method is robust. Two raters, as well as algorithms, each independently find landmarks in template space that are corresponding to the common set of 40 landmarks in Zone 3 in the subject space. Registration accuracy in this zone is measured by the average Euclidean distance between algorithm-calculated corresponding landmarks and rater-labeled corresponding landmarks in the template space. Smaller landmark errors point to a higher registration accuracy in the normal regions far away from recurrent tumors.

4. **Zone 4: the whole brain.** Ideally the boundaries of the warped and the target brain images should be perfectly aligned, since all brain images in this study have been
properly skull-stripped (rater-proofread of the BET-based automatic skull-stripping results). However, in practice, the influence of the recurrent tumors does exist. The extent of the influence directly reflects the robustness of a registration method with regard to the tumor recurrence — less influence means a higher level of robustness in a registration method. Therefore, we measure the dice overlap and the surface distance errors between the warped and the brain masks. A higher dice overlap and smaller surface distance errors are indicative of a higher registration accuracy in aligning brain boundaries, and a higher level of robustness to the tumors.

### 4.4.3. Results and Observations

Figs. 44 and 45, in axial and coronal views respectively, show the representative DRAMMS registration results from a tumor-recurring patient’s brain image to the normal brain template image. The visual results outline the effect of mutual-saliency components in DRAMMS to handle the challenges from missing correspondences. As shown in Figs. 44(d) and 45(d), the automatically-computed mutual-saliency map effectively identifies the outlier regions where correspondences are less likely to be established. The outlier regions coincide with the recurrent tumor regions. This is without any segmentation or prior knowledge of the presence and location of tumor recurrence. As a result, the registration is mainly driven by other normal regions, leading to visually plausible results as shown in Figs. 44(c) and 45(c).

Below are quantitative validations in different image regions (zone 1–4). They will reveal the influence of the recurrent tumors in various regions.

**Registration Accuracy in Zone 1**

Fig. 46 shows the dice overlap and average surface distance errors between the warped tumor recurrence regions. When two human experts are independently mapping the recurrent regions into the same normal template image, the dice overlap between their mapped regions
Figure 44: Registration of a brain image with tumor recurrence to a normal brain template by DRAMMS, in axial view. Without segmentation/initialization/prior-knowledge, the automatically-calculated mutual-saliency map (d), defined in target image space, has effectively assigned low weights to those regions that correspond to those outlier regions (pointed out by arrows) in source image (a). This way, the negative impact of outlier regions is largely reduced; registration is mainly guided by regions that can establish good correspondences. Red arrows point to the post-surgery blood pool. Blue arrows point to the recurrent tumors.
Figure 45: Registration of a brain image with tumor recurrence to a normal brain template by DRAMMS, in coronal view. The caption is the same as in Fig. 44.
is only averaged at around 0.6, reflecting a moderate agreement between human raters. ANTs agrees with raters at the highest level among all 4 methods (average dice 0.56), and is almost comparable to the level that raters agree with each other. DRAMMS follows closely at an average dice of 0.52.

ANTs and DRAMMS obtain similar results in Zone 1, but by quite different registration mechanisms. ANTs treats all image voxels equally, so the recurrent tumor regions do contribute to registration. DRAMMS, on the other hand, treats voxels differently by the automatically-calculated mutual-saliency map. As Figs. 44 and 45 show, the recurrent tumor regions have less role in guiding the registration. In other words, the alignment in tumor recurrence regions is guided by tumor recurrence itself in ANTs, but smoothly interpolated in DRAMMS. Later results in this section, especially results in zone 4 (alignment of the whole image), will show that they have different effects to surrounding regions, and even to the entire image.

Registration Accuracy in Zone 2

Fig. 47 presents landmark errors in the regions close to (< 30mm distance to) the tumor recurrence regions. On average, raters have 4.09mm errors in locating landmark correspondences. DRAMMS has the smallest landmark errors among the 4 methods, followed closely by ANTs and Demons. The errors between these three methods and the raters are comparable to inter-rater errors. So it shows that DRAMMS, ANTs and Demons can perform almost as well as raters in registering the immediate neighbors around tumor recurrence regions.

Registration Accuracy in Zone 3

Fig. 48 shows landmark errors in regions further away from tumor recurrence. Inter-rater error in this zone is 5.33mm, which is surprisingly higher than the inter-rater errors in the close neighborhood of tumor recurrence regions (4.09mm). The errors from algorithms are also higher in regions far away from tumor recurrence than in regions close by. The reason
Figure 46: Registration accuracy in Zone 1 (within the recurrent tumor regions).
might be that, in regions close to tumor recurrence, raters tend to be more conservative to pick up landmarks whose correspondences are more obvious from the other image; therefore raters agree more. The rank of registration methods in Zone 3 is the same as in Zone 2: DRAMMS and ANTs have the smallest errors and performs closest to human raters, followed closely by Demons.

**Registration Accuracy in Zone 4**

Fig. 49 measures the alignment between the warped and the template brain masks. Ideally, brain masks should be perfectly aligned since both are appropriately skull-stripped. However, Fig. 49 shows a big difference between registration methods. DRAMMS has, on average, 1.8 mm error between the warped and the template brain masks, or 1-2 voxel errors (voxel size 1.0 x 1.0 x 1.0 mm$^3$), and the 95-percentile boundary distance is 3.98 ± 0.70 mm in 8 subjects. Demons has slightly bigger boundary distance, with average of 2.15 mm and 95-percentile at 4.41 ± 0.58 mm. Starting from ANTs, the average boundary distance increases to 3.29 mm, and the 95-percentile jumps to 8.69 ± 1.09 mm. The jump in the 95-percentile of boundary distance is observed in every subject by ANTs, generally in the section of the brain boundary that is closer to the recurrent tumors than other boundary sections.
Figure 48: Registration accuracy in Zone 3 (regions far away from tumor recurrence).

shows the global influence of tumor recurrence. In fnirt registration, the global influence of tumor recurrence is more obvious, with the average boundary distance increased to 4.54\( \text{mm} \), and especially with the 95-percentile boundary distance increased to 19.56 \(\pm\) 2.05\( \text{mm} \). This large 95-percentile boundary distance basically indicates that fnirt registration is off in many sections of the boundary, because of the global influence of tumor recurrence regions. The dice overlap between the warped and the template brain mask shows the same trend — DRAMMS and Demons presenting a good alignment in brain masks, indicating less global influence by the recurrent tumors; ANTs is more influenced; and fnirt almost fails.

4.4.4. Conclusion

This section validates the performance of the DRAMMS method in registering brain images with tumor recurrence to the normal template. The main challenge in this registration is the missing correspondence problem caused by the recurrent tumors. We have collected multi-rater annotations of ROI and landmarks to compare both inter-rater and algorithm-to-rater errors. We have also hierarchically studied registration accuracies in different parts (zones) of the images, from within, close to, and far away from the recurrent tumor regions,
Figure 49: Registration accuracy in Zone 4 (the whole image and its boundary).
as well as the whole brain mask.

Combining the results in all the image regions, DRAMMS has the second highest overlap in tumor recurrence regions, the smallest landmark errors in regions close to and far away from tumor recurrence regions, all comparable to human raters, and has the highest accuracy in aligning brain boundaries. This is mainly because of the mutual-saliency component that is flexible to handle different regions and is effective to identify regions where tumor recurred (Figs. 44 and 45).

Compared to DRAMMS, other methods show less robustness to the tumor-induced missing correspondences. Demons is less influenced by the recurrent tumors in far-away regions (Fig. 48) and in brain boundaries (Fig. 49), but presents larger errors within tumor recurrence regions. This is probably due to the transformation model in Demons, where voxels or regions move almost independently and have only local support. In such cases, regions far away from tumor recurrence are less influenced. On the contrary, ANTs scores the highest accuracy in tumor recurrence regions, but suffers from larger errors in aligning the sections of brain boundary that is close to the recurrent tumors. Among the four methods compared in this section, fnirt is most influenced by tumor presence.

In future studies, we will include validations against other registration methods that are specifically developed for registering tumor images. One approach is the cost-function-masking (CFM) strategy. CFM requires manually delineation of the recurrent tumors as well as post-surgery blood pools. Since DRAMMS performs almost as well as human experts in this study, we can expect DRAMMS to be comparable to those registration methods specifically designed for tumor registrations (like CFM approaches).

In conclusion, the performance of DRAMMS is comparable to that of human raters. DRAMMS shows robustness towards the recurrent tumors in almost all regions of the image. ANTs performs similarly, and even slightly better in tumor recurrence regions, but suffers from large brain boundary errors that show the global influence of tumor recurrence. The vali-
dation in this section lays the foundations for applying DRAMMS in the population study of brain tumor patients.

4.4.5. Acknowledgement

This section is a collaborative study with Drs. Michel Bilello and Hamed Akbari, who helped annotate tumor recurrence ROIs and anatomical landmarks throughout 8 pairs of source and target images, and Mr. Xiao Da, who set up and assisted the whole expert annotation process. They have also contributed to the design and discussion in this study. Their contributions are greatly appreciated.
4.5. In Registering Cardiac MR Images Across Subjects

4.5.1. Introduction

Cross-subject image registration rests in the core of many cardiac studies. Examples include atlas constructions (Perperidis et al., 2005), atlas-based segmentations (Zhuang et al., 2010), and morphologic studies to understand disease patterns (Ye et al., 2011).

In the literature, cross-subject cardiac image registration is often handled by voxel-wise registration methods (Makela et al., 2002). Voxel-wise registration methods usually rely on image information only, and do not require anatomic information or human intervention. Therefore, they can be applied to various organs including the heart (Makela et al., 2002). Some basic question remains, however: 1) which voxel-wise registration methods are more accurate and more stable in the cross-subject cardiac registration context; 2) whether those more accurate methods in cardiac registration coincide with those in brain image registrations (e.g., as found in (Klein et al., 2009)). The answers to these questions are not immediately clear, largely because the heart is usually imaged with a lower resolution, lower signal-to-noise ratio (SNR), more severe moving artifacts, and has a very different shape than the brain.

Towards answering these questions, this section evaluates 12 publically-available registration methods and validates DRAMMS (Ou et al., 2011) in the context of cross-subject cardiac registrations.

We have collected short-axis end-diastole magnetic resonance (MR) images of 24 subjects. By permuting the source and target images, this dataset results in 552 possible pair-wise registrations for each registration method. The large number of experiments (perhaps largest to date in the cardiac registration context) is the first feature of this study. The second feature of this study is the comprehensive evaluation criteria. Unlike other evaluation
studies (e.g. [Klein et al., 2009]) that only measure accuracy, we measure both accuracy and aggressiveness of deformations, and visualize their relationship in a joint plot. A deformation is considered more "aggressive" if it leads to self-foldings at more locations, and if it takes greater expansions/shrinkages to capture cross-individual variations. Aggressiveness and accuracy are usually a pair of trade-off. A higher accuracy often comes from an increased level of aggressiveness in deformation. On the other hand, an over-aggressive deformation can undesirably break topology. An ideal method should achieve a high registration accuracy while preserving topology. Measuring both accuracy and aggressiveness helps reveal which methods better balance the two. The third feature of this study is that, instead of using only one set of parameters, we have examined two parameter settings for the four most accurate methods – one set of parameters corresponding to more aggressive deformations and one set corresponding to smoother deformations. This is important, because different cardiac studies have different requirements on the aggressiveness levels of deformation. It also helps reveal which methods achieve a consistently high accuracy when aggressiveness levels change. Results shown in this section have been published in a recent proceeding paper [Ou et al., 2012].

In the rest of the section, we present the evaluation protocol in Section 4.5.2 and the evaluation results in Section 4.5.3. We discuss and conclude this section in Section 4.5.4.

4.5.2. Evaluation Protocol

This subsection describes our evaluation protocol. It contains three parts: the datasets (Section 4.5.2.1); the registration methods included (Section 4.5.2.2); and the evaluation criteria (Section 4.5.2.3).
4.5.2.1. Dataset for Evaluation

We now describe the dataset and pre-processing steps. Three-dimensional short-axis cardiac MR images of 24 subjects are collected at the end-diastole (ED) phase. The image dimension is $120 \times 120 \times 12$ and the voxel size is $1.25 \times 1.25 \times 8.0 \text{mm}^3$. Common pre-processing steps include the correction of respiratory motions (Zhang et al., 2010) and the N3-based bias field correction (Sled et al., 1998). Non-cardiac structures are removed by a semi-automatic process. In this process, the heart is first automatically outlined by a public software “Segment” (Heiberg et al., 2010). Then, a cardiovascular expert refined the separation of cardiac and non-cardiac structures. Removal of non-cardiac structures is similar to skull-stripping in brain image registrations. The purpose is to remove unnecessary challenges, especially when different images may contain different non-cardiac structures due to different fields of view. Each cardiac image is further annotated by the same cardiologist into three structures – left ventricle (LV), right ventricle (RV) and myocardium. Some typical intensity images and expert-annotation images are shown in Fig. 50. We note that, except for removing non-cardiac structures, these expert annotations of LV/RV/myocardium are in no means used as any part of the registration process. They are only used to evaluate registration accuracy.

This dataset represents the common challenges in cardiac registrations – lower resolution, lower SNR, more severe moving artifacts and quite different shape from the brain. Besides, 11 out of 24 subjects have tetralogy-of-fallot (TOF) defect, and hence having irregular ventricle shapes largely different from the remaining 13 normal subjects (Fig. 50).

4.5.2.2. Registration Methods to be Evaluated

The same set of 12 publicly-available methods from the brain study (Section 4.1.2.2 and Table 4) are included here for cardiac registration evaluation. More details of the 12 registration methods can be found in the literature review chapter of this dissertation. The
Figure 50: Images (a) and expert-annotation of structures (b) for some 10 typical subjects from the dataset used in this study. Subjects in the first row in (a) are healthy controls and in the second row are with tetralogy-of-fallot (TOF) defect. In the expert annotation, white, orange and blue regions are LV, RV and myocardium, respectively.
reasons for choosing them are the same too, for the wide variety of registration approaches they represent. That is, they have different similarity measures, different deformation models and different optimization strategies, which are the most important components for registration algorithms.

Unlike our brain studies, where the optimized parameters of each software package is provided by the authors themselves, here in cardiac images we need to choose a set of parameters for each software package. Different parameters can lead to different performances — either a higher accuracy at more aggressive deformations, or reverse. We set parameters by the following two rules. First, we tune parameters not just for the best accuracy, but for the best accuracy at a similar aggressiveness level. Specifically, we start from parameters in a method’s user manual or papers in the past. In each iteration, we keep other methods’ parameters fixed, and slightly adjust one method’s parameters until its deformations are at similar levels with most other methods (few or no self-foldings, similar min, max and range of Jacobian determinants). We iterate on every method until they all converge to similar aggressiveness level. This provides a common ground for more objectively evaluating their accuracies. Second, we provide two sets of parameters, instead of only one set corresponding to the highest accuracy, for the four most accurate methods. One set of parameters correspond to an generally higher accuracy but the increased risk of self-folding; and another set of parameters correspond to generally smoother deformation but at a lower accuracy. This reveals the stability of a registration method as parameters change.

To avoid the bias in selection a target image, we have considered all possible images as the source and the target images in registration. This results in 552 (= 24 × 23) possible pair-wise registrations for each registration method.

4.5.2.3. Evaluation Criteria

Similar to the evaluation criteria in the brain study (Section 4.1.2.3), we measure both the deformation accuracy and the deformation aggressiveness in the cross-subject registration
of cardiac images.

Specifically, deformation accuracy is measured by the Jaccard Overlap \cite{Jaccard1901} between the deformed expert-annotation of the source image and the expert-annotation of the target image. We measure overlaps in 3 regions: LV, RV, and myocardium. A larger overlap often indicates a greater spatial alignment between subjects \cite{Klein2009,Christensen2006}.

Deformation aggressiveness is measured by 4 Jacobian-based metrics: 1) the number of deformations having negative Jacobian determinants; 2) the percentage of voxels having negative Jacobian determinants; 3) the minimum and 4) the maximum Jacobian determinants in a deformation. Finally, we use one metric, the range of Jacobian determinants (=maxJac-minJac), to quantify the aggressiveness of a deformation. ITK calculator is used to compute Jacobians of deformation, for the same reasons described in Section \ref{sec:methods_deformation}.

4.5.3. Results and Observations

We now present the evaluation results (accuracy, aggressiveness, and their correlation) in this section. Observations follow each set of results.

A. Deformation Accuracy

The Jaccard overlap-indicated deformation accuracy is shown in Fig. [51] for myocardium, LV and RV structures. Several observations can be made:

A1) in general, voxel-wise registration methods evaluated in this section have obtained a 0.6-0.9 Jaccard (roughly 0.75-0.95 Dice) overlap in the left and right ventricles, and a 0.4-0.7 Jaccard (roughly 0.55-0.85 Dice) overlap in the myocardium.

A2) DRAMMS scores the highest Jaccard overlap in all three structures in this dataset – an average 0.85 Jaccard (0.9 Dice) in LV and RV, and 0.7 Jaccard (0.8 Dice) in myocardium. The margin is bigger in myocardium regions. A plausible explanation is that DRAMMS
Figure 51: Box-and-Whisker plots: registration accuracy indicated by the Jaccard overlap in 3 expert-annotated structures. From top to bottom, results for myocardium, LV and RV regions. Letter "A" stands for aggressive version and "S" for smooth version of a method.
uses texture attributes other than solely intensity information to define the similarity at each voxel.

**A3)** ANTs, MI-FFD, Demons, and ART also obtained high overlaps in this cardiac dataset. This echoes the findings in a brain registration evaluation study \cite{Klein2009}.

**A4)** Methods using intensity differences (SSD) as similarity metrics have a reasonable Jaccard overlap on average. However, they have larger variations, and suffer in difficult cases. This shows that SSD metric is less likely to consistently capture large anatomical variations. One solution is to combine the intensity difference with deformation mechanism of more degrees of freedom (like in ART and Demons). A perhaps better solution is to replace it with a more robust similarity metric, such as the correlation (like in ANTs), mutual information (like in MI-FFD), or an attribute-based similarity (like in DRAMMS).
B. Deformation Aggressiveness

The four sets of Jacobian statistics indicating deformation aggressiveness are shown in Fig. 52. From left to right, top to bottom, they are: the number of deformations with negative Jacobian determinants; the percentage of voxels having negative Jacobian determinants; the minimum and the maximum Jacobian determinants in deformations. We observe the following from the results:

B1) From the top row in Fig. 52, fnirt is the only non-rigid registration method that guarantees diffeomorphism in this dataset. Diffeomorphism means no existence of negative Jacobian determinants (i.e. no self-folding) in deformations. It is a nice property that preserves topology and one-to-one forward and backward correspondences. fnirt guarantees diffeomorphism by directly checking and removing negativities in the Jacobian map. However, this is at the cost of the overlap-indicated registration accuracy, as reflected in Fig. 51. Actually, whether cross-subject deformation is a diffeomorphism is an unknown matter, especially when there are large anatomic variations.

B2) DRAMMS(A), ANTs(A), MI-FFD, Demons and ART(A) score the higher overlap in Fig. 51. Interestingly, results in the lower row of Fig. 52 show that they have quite different deformation styles. In particular, DRAMMS(A), ANTs(A) and MI-FFD have greater maximum Jacobian determinants, trying to capture the individual variability with larger expansions. Demons and ART(A) have more negative minimum Jacobian determinants, trying to capture the same individual variability with more self-foldings in deformations.

C. Correlation between accuracy and aggressiveness surrogates

The surrogates for deformation accuracy (Jaccard overlap) and the surrogates for deformation aggressiveness (Jacobian range) are plotted jointly in Fig. 53. Here the y-axis is the mean Jaccard overlap over all 3 structures and all 552 registrations, indicating the overall accuracy of a registration method. The x-axis is the mean range of Jacobian determinants (=mean(maxJacobianDet-minJacobianDet)) over all 552 registrations, indicating the aggressiveness level of a method. Three observations can be made from this figure:
C1) Methods score higher overlaps at more aggressive deformations, which is expected.

C2) An ideal registration method should obtain a highest possible overlap while preserving topology (being diffeomorphism). Combining Fig. 53 with the upper left part of Fig. 52, DRAMMS(S), the smooth version of DRAMMS, obtain the second highest overlap and the diffeomorphic deformations in almost all but 3% (17/552) deformations.

C3) In Fig. 53 we use dashed lines to connect the smooth and aggressive versions of four top-ranking methods. As a result, we observe that DRAMMS, in general, have high accuracies. More importantly, it has a greater increase when going from the smooth to the aggressive versions. It therefore offers a wider range of choices for varying needs. That is, the aggressive version, DRAMMS(A), seems to be a good candidate for single-/multi-atlas-based segmentation, where the overlap is the focus. The smoother version, DRAMMS(S), is perhaps a better choice for finding the common disease pattern in a population, where the key is to remove the global difference among images and meanwhile preserve the disease-induced individual variability.

4.5.4. Discussion

This section validates DRAMMS among 11 other voxel-wise registration methods within the context of cross-subject cardiac registrations in a dataset of 24 subjects. Results show that those top-ranking registration methods – DRAMMS, ANTs, MI-FFD, Demons, ART – obtain an average Jaccard overlap of 0.7-0.9 (i.e. Dice of 0.82-0.95) in the left and right ventricles, and 0.5-0.7 (i.e., Dice of 0.66-0.82) in the myocardium. In the following, we discuss the important aspects in this section.

The objectivity in the evaluation is a critical issue. In our study, it is encouraged by looking at accuracies when most methods are at similar aggressiveness levels. Deformation accuracy
and aggressiveness are often a pair of trade-off. Reporting both and correlating them are a more comprehensive set of criteria than purely the accuracy criterion. The accuracy and aggressiveness results (Figs. 51, 52), and their correlation (Fig. 53), show that the smooth version of DRAMMS achieves best balance in this dataset – a high overlap and the maximum preservation of diffeomorphism. ANTs, MI-FFD, Demons and ART also perform well in this cardiac dataset. This echoes the findings in brain registration study (Klein et al., 2009).

On the note of similarity metrics, the intensity difference metrics are often less stable than the correlation (like in ANTs), mutual-information (like in MI-FFD) or attribute-based similarity (like in DRAMMS). On transformation models, different behaviors are observed. The deformation models behind DRAMMS, ANTs and MI-FFD tend to capture the individual variability by larger deformation expansions and less severe self-foldings. Models behind Demons and ART tend to behave to the opposite.

One surprising observation is regarding diffeomorphism. fnirt is the only one that guarantees
diffeomorphism in this dataset, as it directly checks and removes negative Jacobian determinants. Non-diffeomorphism occurs for many methods, although some were theoretically designed diffeomorphic. Numerical issues might be one reason. Or, perhaps the process of deforming subjects with large anatomical variability itself is not completely diffeomorphic in nature. We note that the Jacobian calculator would play a role, too.

Future work includes additional validations that consist of additional registration methods, cardiac datasets, and accuracy surrogates like the surface distance.
4.6. In Registering Breast Cancer Images Over Time

4.6.1. Introduction

This section validates the performance of DRAMMS in the registration of longitudinal breast images. In radiation oncology, breast cancer patients undergoing chemotherapy are usually followed up with periodic breast MR scans to track treatment effects. The routine clinical practice is to have radiologists outline breast tumors in images from different time points, and do a visual registration to see how breast tumors have changed. Besides being subjective and time-consuming, the visual registration often captures global changes in the entire tumor region. In this domain, automated longitudinal image registration methods are expected to help, with a higher level of objectivity, reproducibility. More importantly, automatic registration is capable of capturing more local changes (Arlinghaus et al., 2010). This is important, as tumor indeed changes both globally and locally as the response to chemotherapy.

The breast mainly consists of soft tissue (parenchyma and fat). Compared to other organs like the brain, the heart or the prostate, the breast presents higher deformations over time. This is further complicated by the differences in imaging positions during breast MRI, and the longitudinal breast tumor change in size, shape and textures as the response to chemotherapy. These factors render longitudinal breast registration a fairly challenging task.

Various breast registration methods have been developed in the literature (Rueckert et al., 1999; Rohlfing et al., 2003; Li et al., 2009b, 2010). They are mostly based on general-purpose registration methods. Physical constraints specifically designed for breast registration can be added into the registration pipeline, such as the incompressibility constraints (Rohlfing et al., 2003). The validations of these methods are often on simulated data, or by comparison...
In this section, we have landmark definitions from two raters. This enables us to observe
inter-rater differences as well as algorithm-to-rater differences. A registration algorithm is
said to be as good as human experts if its average registration errors with raters are com-
parable or equivalent to the inter-rater errors. Therefore, a more comprehensive validation
with multi-rater data is the first feature of the work presented in this section. Furthermore,
as most existing studies only report overall registration errors, the experiments in
this section report registration errors in stratifications: 1) the registration errors in those
who respond well to chemotherapy (i.e., the complete disappearance of breast tumors at the
end of chemotherapy) and those who do not respond well to chemotherapy (detailed defini-
tion in 4.6.2.1); and 2) registration errors in normal regions, tumor regions and the entire
image. The patient stratification and the region subdivision are important features in this
study. They promote further investigations or prediction of the response to chemotherapy,
which is crucial for clinical decision-making.

In the following, Section 4.6.2 introduces the validation protocol (data, registration methods
included, and evaluation criteria); Section 4.6.3 presents the validation results; and Section
4.6.4 concludes this section.

4.6.2. Validation Protocol

This subsection describes below 1) the datasets; 2) the registration methods to be evaluated;
and 3) the evaluation criteria.

4.6.2.1. Datasets for Evaluation

Fourteen breast cancer patients undergoing neoadjuvant chemotherapy are included in this
study. They are patients recruited for the I-SPY clinical trial (Investigation of Serial Studies
to Predict Your Therapeutic Response with Imaging And Molecular Analysis). They have been followed-up at bi-weekly or monthly intervals with T1-weighted MRI during the course of chemotherapy. At the end of the 3-4 months of chemotherapy, patients are evaluated by pathology analysis. Those who show the absence of any residual invasive cancer in the breast and the absence of any metastatic cells in the regional lymph nodes are defined as "responders", or in more clinical terms, pathologic complete response (pCR) (Chen et al., 2008; von Minckwitz et al., 2012). The remaining patients are defined as "non-responders", or patients with incomplete response. By this definition, 6 patients are classified as non-responders and 8 as responders. Evaluating registration accuracies separately in responders and in non-responders is an important feature of our evaluation study. Responders show larger tumor changes compared to non-responders, therefore should be more difficult to register over time.

All images are T1-weighted MR images. Image sizes and voxel sizes slightly vary among subjects, but a typical image size is $256 \times 256 \times 64$ and a typical voxel size is $0.78 \times 0.78 \times 2.0mm^3$. We register the second follow-up image to the baseline image for each patient. The baseline image is the image obtained right before the chemotherapy. The second follow-up image is taken $31 \pm 21$ days after baseline image in this dataset. Follow-up images further away from the baseline images should present larger deformations, and hence the increased difficulty for registration. We choose the second follow-up image out of all 3 or 4 available follow-ups for each patient, such that the registration of follow-up images to the baseline image is in the median level of difficulty. In the future study, it is of great interest to evaluate the registration of all follow-up images into the baseline image space. This can further reveal the accuracy of registration at different times.

Two breast cancer experts, Drs. Emily Conant and Susan Weinstein, both from the Department of Radiology at the University of Pennsylvania, defined landmarks independently in the images. First, one doctor (SW) defined corresponding landmark pairs between the baseline and the follow-up images for each patient ($85 \pm 29$ landmark pairs for each patient).
Table 5: Lists of patients in the responder and non-responder groups. The time elapse between the follow-up and baseline images, and the number of landmarks in tumor and normal regions are listed for each patient. (FU: follow-up; BL: baseline; LM: landmarks.)

<table>
<thead>
<tr>
<th>Non-Responders</th>
<th>#days(FU-BL)</th>
<th>#LM (tumor)</th>
<th>#LM (normal)</th>
<th>#LM (all)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BL</td>
<td>14</td>
<td>63</td>
<td>59</td>
<td>122</td>
</tr>
<tr>
<td>DP</td>
<td>29</td>
<td>25</td>
<td>55</td>
<td>80</td>
</tr>
<tr>
<td>HR</td>
<td>13</td>
<td>64</td>
<td>60</td>
<td>124</td>
</tr>
<tr>
<td>KB</td>
<td>14</td>
<td>22</td>
<td>45</td>
<td>67</td>
</tr>
<tr>
<td>ME</td>
<td>14</td>
<td>90</td>
<td>16</td>
<td>106</td>
</tr>
<tr>
<td>MK</td>
<td>56</td>
<td>53</td>
<td>55</td>
<td>108</td>
</tr>
<tr>
<td>mean ± stdev</td>
<td>23 ± 17</td>
<td>53 ± 26</td>
<td>48 ± 17</td>
<td>101 ± 23</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Responders</th>
<th>#days(FU-BL)</th>
<th>#LM (tumor)</th>
<th>#LM (normal)</th>
<th>#LM (all)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE</td>
<td>61</td>
<td>23</td>
<td>54</td>
<td>77</td>
</tr>
<tr>
<td>GV</td>
<td>14</td>
<td>14</td>
<td>48</td>
<td>62</td>
</tr>
<tr>
<td>KK</td>
<td>60</td>
<td>16</td>
<td>45</td>
<td>61</td>
</tr>
<tr>
<td>MB</td>
<td>56</td>
<td>8</td>
<td>19</td>
<td>27</td>
</tr>
<tr>
<td>MD</td>
<td>14</td>
<td>22</td>
<td>81</td>
<td>103</td>
</tr>
<tr>
<td>MI</td>
<td>55</td>
<td>41</td>
<td>20</td>
<td>61</td>
</tr>
<tr>
<td>RC</td>
<td>21</td>
<td>51</td>
<td>25</td>
<td>76</td>
</tr>
<tr>
<td>SD</td>
<td>14</td>
<td>31</td>
<td>8</td>
<td>79</td>
</tr>
<tr>
<td>mean ± stdev</td>
<td>37 ± 23</td>
<td>26 ± 14</td>
<td>47 ± 26</td>
<td>73 ± 28</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>All Subjects</th>
<th>#days(FU-BL)</th>
<th>#LM (tumor)</th>
<th>#LM (normal)</th>
<th>#LM (all)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean ± stdev</td>
<td>31 ± 21</td>
<td>37 ± 24</td>
<td>48 ± 22</td>
<td>85 ± 29</td>
</tr>
</tbody>
</table>

Then, the second doctor (EC) independently defined landmarks in the follow-up images that, according to her expert knowledge, are corresponding to the same set of landmarks in the baseline image. These landmarks are further labeled as being within tumor regions or normal regions. This leads to registration evaluations in different regions.

Table 5 lists more detailed information of each patient — responder or non-responders, the time elapse between follow-up and baseline images; and the number of landmarks in tumor, normal and all regions.
4.6.2.2. Registration Methods to be Evaluated

DRAMMS is compared with two raters. The inter-rater errors and DRAMMS-to-rater errors show whether DRAMMS performs as well as human experts.

Besides, two other methods are compared: flirt, an affine registration method, and ART, one of the most accurate deformable registration algorithms in previously-evaluated brain registration [Klein et al., 2009]. More detail of these methods can be found in the previous Section 4.1.2.2 and in the literature review chapter of this dissertation.

The same to the experiments in other sections of this chapter, parameters remain the same for each method. They are not necessarily optimal for this specific longitudinal breast registration task. However, such comparison still shows how each registration method, with its fixed set of parameters, performs in various registration tasks. And that is how an ordinary user will use those registration methods in most cases.

4.6.2.3. Evaluation Criteria

For each patient, raters and algorithms map the same set of landmarks from the baseline image into the follow-up image space. Therefore, landmark errors (i.e., the Euclidean distance between the mapped landmark locations) are all measured in the follow-up image space. The algorithm-to-rater error is the average error between the algorithm and the two raters (i.e., \(=(\text{Error(algorithm v.s. Rater1)}+\text{Error(algorithm v.s. Rater2)})/2\)). Smaller landmark errors indicate a higher registration accuracy. Also, an algorithm performs more similarly to human experts if its algorithm-to-rater errors are closer to inter-rater errors.
4.6.3. Results and Observations

Fig. 54 shows a typical set of longitudinal registration results by DRAMMS. Although breast MR images usually present high noise levels, and large changes over time in both tumors and normal soft tissue, DRAMMS has obtained visually good alignment in various areas of the image.

For the quantitative results comparing with raters and other registration methods, Fig. 55 shows the overall errors averaged among all subjects. Two experts have an average disagreement of 3.63mm for finding landmark correspondence locations. DRAMMS has the smallest overall error among automatic registration methods compared here and is closest to human experts.

Fig. 56 shows landmark errors in patient sub-groups (responders v.s. non-responders), and in different image regions (tumor, normal and overall). Several observations can be made:

a) Registration errors in responders are bigger than in non-responders. This is the trend for both raters and algorithms, and in all image regions. We can probably attribute this to the larger deformation and larger tumor change in responders.

b) Responders and non-responders share similar registration errors in the normal regions of the breast.

c) Responders and non-responders have different registration errors in the tumor regions. The errors in tumor regions are bigger than the errors in normal regions for responders; whereas it is the reverse in non-responders. This agrees with the fact that tumors change more in responders than in non-responders in response to chemotherapy.

c) Overall we should expect about 4mm registration errors throughout the whole breast image, either by human experts or by the most accurate registration algorithms.
Figure 54: Results for longitudinal registration of breast cancer images by DRAMMS.
4.6.4. Conclusion

This section evaluates the performance of DRAMMS in the longitudinal registration of breast cancer MR images. Longitudinal registration is the basis to automatically quantify the breast cancer change over time. It is usually complicated by differences in positioning of the breasts, by large deformations of the soft tissue over time, and by large changes of breast tumors in size, shape and texture during the course of chemotherapy. This section shows that the performance of DRAMMS is comparable to that of human experts, and better than the flirt and ART methods.

Multi-rater landmarking is one important feature in this study. Ours is perhaps the first such study in the literature. With this experiment design, we show that the breast registration is expected to have about 4mm correspondence errors even by human experts. This error range, clearly larger than errors in the brain and cardiac registrations, underlines the difficulty of breast registration over time. Dividing patients into subgroups and dividing images into tumor and normal regions mark another important feature of this section. Together they show that the difficulties of registering normal regions are almost the same in responders and non-responders; however it is more difficult to register tumors for respon-
Figure 56: Accuracy in longitudinal breast cancer image registrations. Accuracies are reported in responders and non-responders, and in tumor, normal and all regions.

Responders, making it more difficult to register the whole image for responders. The reason should be the larger tumor change in responders due to the response to chemotherapy.

Future work is on validating registrations in images from more follow-up times. It will reveal how registration accuracy changes when the two images to be registered are taken from time points further away, and how such changes differ in responders versus non-responders. The change pattern (e.g. tumor change rate) might be used as indicators for the differentiation of responders and non-responders.

4.6.5. Acknowledgement

This section is a collaborative study with Drs. Emily Conant and Susan Weinstein, who helped label anatomical landmarks throughout the 14 pairs of source and target images, Dr. Despina Kontos, who helped set up collaborations with doctors and contributed to the design and discussion of this study, and Dr. Sarah Englander, who helped develop graphic user’s interface (GUI) for doctors’ landmark labeling. Their contributions are greatly ap-
preciated.
4.7. Conclusions

This chapter presents extensive validations of the DRAMMS algorithm developed in Chapter 3. Validations are in brain, cardiac and breast images, and cover the following registration tasks:

1. Cross-subject Registration of Skull-stripped Brain MR Images;
2. Cross-subject Registration of Raw Brain MR Images;
3. Registration of Lesion-bearing Brain MR Images with a Normal Brain Template;
4. Registration of Tumor-recurring Brain MR Images with a Normal Brain Template;
5. Cross-subject Registration of Cardiac MR Images; and

In each registration task, DRAMMS is quantitatively compared with other state-of-the-art registration methods, and evaluated against expert-defined structure and/or landmark annotations in public as well as in-house datasets.

The results in this chapter demonstrate the three properties of DRAMMS: a) generally applicable to different organs (brain, heart, breast) and different registration settings (cross-subject, longitudinal); b) accurate compared to the state-of-the-art top accurate registration methods and compared to human experts; c) robust with regard to intensity inhomogeneities, contrast difference, cross-subject anatomical variations, longitudinal structure changes, and even with regard to partial missing correspondences that may be caused by lesions, tumors or the difference in imaging field-of-view (FOV).

Despite extensive validations, we will note several situations in which DRAMMS does not apply or may fail. 2D-to-3D registration, such as registering a 2D X-ray image to a 3D CT or
MR images, is one scenario that DRAMMS does not apply. Group-wise registration, where images from a group of subjects are registered together into a virtual or actual population mean image space, is another scenario where DRAMMS does not apply. Even in 3D pairwise registrations, it is not rare to find DRAMMS fail in very challenging tasks. For example, DRAMMS may often fail when registering raw cardiac images between different subjects. This is because the large variation in different subjects’ hearts, the considerably different FOVs, and especially because the surrounding non-cardiac structures, which dominant the image space, differ greatly among subjects. That is the reason that DRAMMS is only validated in the registration of pure cardiac images in Section 4.5. DRAMMS may also fail when registering raw prostate images. This is because of the complications in the non-prostate structures (e.g., the bladder, the rectum), which dominant the image space.

In conclusion, this chapter demonstrates the general applicability, accuracy and robustness of the proposed DRAMMS method in various registration tasks involving brain, cardiac and breast images. The extensive validations in this chapter lay the foundations to applying DRAMMS into a wide variety of clinical and research studies, including population study, longitudinal study, and multi-atlas-based segmentation. The next chapter will present five examples.
CHAPTER 5

Applications of DRAMMS

To show the wide application of the DRAMMS registration algorithm, this chapter applies
DRAMMS to the five clinical and research studies. They are examples of population studies,
longitudinal studies and atlas-based segmentations. Below is a list of them:

1. the construction of statistical atlases for brain lesions in Type II diabetic population
   (Section 5.1);

2. the construction of statistical atlases for tumor recurrences in brain tumor population
   (Section 5.2);

3. the quantification of breast cancer changes during the course of chemotherapy (Section
   5.3); and

4. multi-atlas-based brain skull-stripping (Section 5.5); and

5. multi-atlas-based brain skull-stripping (Section 5.5).

In all these studies, image registration plays a central role. They also pose considerable
challenges to image registration, due to the presence of pathologies (white matter lesions,
tumors), field-of-view differences (especially in multi-site datasets such as in Study 5 above),
and large anatomical variations. In the following, each section presents one study. Finally,
Section 5.6 concludes this chapter.
5.1. Construction of Statistical Atlases for Brain Lesions in Diabetic Population

5.1.1. Introduction

Patients having Type II diabetes are often associated with white matter lesions (WMLs) in the brain (Schmidt et al., 2004; Jongen et al., 2007; Anan et al., 2009). MRI is often used to observe WMLs due to its high sensitivity and contrast in deep brain structures (Anan et al., 2009; Starr et al., 2003). In the Action to Control Cardiovascular Risk in Diabetes (ACCORD) clinical trial (Buse et al., 2007), we have analyzed T1-weighted brain MR images from 487 patients having Type II diabetes.

One important analysis in the ACCORD trial is to derive statistical atlases of white matter lesions (WMLs) in the diabetic population. A statistical atlas is the spatial average of WML loads in this diabetic population. It shows where and how often lesions occur in the population. The spatial location and distribution of these lesions might relate not only to the clinical profile of a population, but might also be predictive of the future cognitive decline and of the response to therapies (e.g., intensive control of blood glucose and blood pressure).

To construct statistical WML atlases, the fundamental step is to register images from all patients into a common template space. The template is often chosen from a normal subject outside the patient population, so as to represent the healthy anatomy. Image registration is, however, often challenged by the lesion-induced missing correspondences. DRAMMS can apply to this situation, as Section 4.3 already show its relatively high level of robustness with regard to the presence of lesions compared with other image registration methods.

The constructed statistical atlases will lead to two important findings:

1. To reveal the differences in average WML loads between different sub-populations
stratified by clinical variables (gender, duration of diabetes and Hb1Ac values). Here Hb1Ac is an index of hemoglobin that indicates the blood glucose level in diabetes patients.

2. To differentiate treatment effects by different treatment strategies. Statistical atlases of the same population before and after treatment will show the change in the average WML loads as the effect of treatments. This provides opportunities to observe different effects by different treatment strategies.

This section constructs statistical atlases for the two purposes above. In the following, Section 5.1.2 describes material and methods for atlas construction; Section 5.1.3 shows and analyzes results; and finally Section 5.1.4 discusses future work and similar studies where DRAMMS can apply.

5.1.2. Material and Methods

In the ACCORD clinical trial, 487 patients are randomly selected to undergo one of two different treatment strategies. One strategy, known as the "standard treatment strategy", aims to lower the glycemia level to 7-7.9%. The other strategy, known as the "intensive treatment strategy", aims to lower the glycemia level even further to 6%. The trial has collected T1-weighted brain MR images for each patient at the beginning of the treatment (baseline) and 40 months into the treatment (follow-up). Patients are stratified into sub-populations by gender (261 males v.s. 226 female), by the duration of diabetes (143, 169, 93, 45, 37 subjects having diabetes less than 5 years, 5-10 years, 11-15 years, 16-20 years and 21-40 years), by H1bAc values (28 subjects with H1bAc<7.0, 239 subjects between 7.1–8.0, 147 subjects between 8.1–9.0, 73 subjects with H1bAc>9.0), and by different treatment strategies (standard v.s. intensive).

Images of all patients are registered into the same template image that has normal anatomy.
DRAMMS is used to register the lesion-bearing T1-weighted MR images into the T1-weighted MR template space. The same deformation calculated by registering T1-weighted intensity images are used to warp lesion regions from the patient space into the template space. The lesion regions in the patient space are segmented by an automatic white matter lesion segmentation tool (Lao et al., 2008). This automatic segmentation tool is chosen because it has achieved a white matter lesion segmentation accuracy comparable to that of the human expert’s in cross-validations. The warped lesion regions from a specific subpopulation are averaged, at each voxel, in the template space. The averaged lesion maps are referred as statistical atlases of white matter lesions in this section.

5.1.3. Results

Atlases are presented in this sub-section to observe 1) the difference in patients with different clinical variables (gender, diabetes duration, H1bAc values), and 2) the difference between different treatment strategies.

A. Statistical WML Atlases for Different Clinical Variables

Fig. 57 shows statistical WML atlases in color-coded form for males and females. Both
atlases are overlaid onto the template image to reveal the normal anatomy around lesion occurrence. Visually male and female Type II diabetic patients have similar chances of developing similar loads of WMLs in similar locations (mostly in peri-ventricle and horn areas). As expected, no clear gender difference is observed in terms of the average WML occurrence.

Fig. 58 shows statistical WML atlases for patients having different years of diabetes. It shows an expected trend that patients are more likely to develop white matter lesions in wider areas as the duration of diabetes increases. Visually the change is gradual over years.

Fig. 59 shows statistical WML atlases for different H1bAc values. H1bAc values reflect blood glucose levels. A patient is diabetic if the H1bAc value is greater than 6.5. A higher H1bAc value indicates an increased level of glucose in the blood, and hence more severe diabetic conditions. As expected, patients with higher H1bAc values have shown higher chances of seeing white matter lesions in peri-ventricle areas.

B. Statistical WML Atlases for Tracking Treatment Effects

Fig. 60 shows the statistical WML atlases of all 487 patients in the baseline and the follow-up time points (40 months later). Patients have undergone either standard or intensive
5.1.4. Discussion

This section applies DRAMMS registration into constructing statistical atlases of brain white matter lesions in the Type II diabetes patients. It is an example of how DRAMMS can be used to normalize images from a population of subjects into a common template space, so that they can be jointly analyzed. The constructed atlases have visualized that Type II diabetic patients are more likely to see white matter lesions in the peri-ventricle and horn areas. The chances of developing brain white matter lesions are higher when
Figure 60: Statistical WML atlases (overlaid on the template image) at baseline and follow-up (40 months) for all 487 patients.

Figure 61: Statistical WML atlases (overlaid on the template image) at both baseline and follow-up (40 months) for patients undergoing the standard treatment strategy and patients undergoing the intensive treatment strategy.
patients have longer time with diabetes, or when patients have higher blood glucose levels as indicated by the HbA1c values. Over years, lesions grow regardless of which treatment strategy is used. And the intensive treatment strategy does not necessarily slow down the lesion growth compared to the standard treatment strategy.
5.2. Generation of Post-Surgery Brain Tumor Recurrence Maps

This section presents preliminary results of applying DRAMMS in the study of tumor population. The framework is very similar to the study the brain lesions in the previous section.

5.2.1. Introduction

Brain tumors are intracranial solid neoplasms that often pose serious life threats. Routine treatments include the surgical resection of tumors, followed by chemotherapy or radiotherapy. Because of its highly invasive and infiltrative nature, brain tumor often recurs after treatment. Knowing where tumor recurs, how the recurrence relates to the initial occurrence, and whether tumor recurrence follows certain patterns (e.g., along major vessels or white matter pathways) will provide valuable information for predicting post-surgery disease progression and adjusting treatment strategies prior to the surgery.

One straightforward way of answering the above questions is to build statistical atlases of both the initial and the recurrent tumors of the same population. In order to have the atlases of the initial tumor occurrence and the atlases of tumor recurrence comparable, they should both reside in the same normal template space. Ideally, blood vessel images (such as MR Angiography) or white matter connectivity maps should also be available in the same template space, so that the recurrence patterns can be better interpreted.

Similar to constructing white matter lesion atlases in Section 5.1, the central piece in constructing tumor recurrence atlases is the registration of images from a population of brain tumor patients into a common template space. DRAMMS applies here, as it has already
been validated in Section 4.4 to achieve a comparable accuracy with that of human experts' and the highest accuracy among competing methods.

5.2.2. Materials and Methods

In our experiment, we have collected 11 patients having recurrent tumors following the surgical resection of the initial tumors. Their recurrent tumor regions have been delineated by neuro-radiologists in images from a specific MRI modality known as "fluid attenuated inversion recovery" (FLAIR). Their T1-weighted MR images are also collected. By registering their T1-weighted images onto the same template image using DRAMMS method, their expert-delineated tumor recurrences are warped into the template space. The warped tumor recurrence regions from all subjects are averaged at the voxel level to form a statistical atlas of tumor recurrence in this population.

Here, FLAIR image is used to label tumor recurrence because of its special pulse sequence design to contrast tumor and surrounding areas (Tsuchiya et al., 1996; Chen et al., 2001). On the other hand, T1-weighted images are more appropriate for registration because of the high contrast between tissue types and in deep brain structures. A patient’s FLAIR and T1-weighted images are naturally aligned, so the tumor delineation in FLAIR images is automatically tumor delineation in T1-weighted images, which can be mapped into template space by patient-to-template registration based on T1-weighted images.

5.2.3. Results

Fig. 62 shows the constructed tumor recurrence atlas overlaid on the normal template image. The atlas at this stage offers no clear observations, since it is only constructed from a very small number (11) of patients, and these patients have initial tumor occurrence in different regions. Ideally, we will recruit tens or hundreds of patients having initial tumors

171
at roughly the same regions. That way, we can specifically observe the patterns of tumor recurrence for the initial tumors at some specific regions.

5.2.4. Discussion

This section applies DRAMMS into constructing tumor recurrence atlases. The purpose is to study tumor recurrence patterns, e.g., the distance and recurrence paths from the initial tumor locations. The currently constructed atlas does not offer clear observations due to the small sample size (11 subjects). However, the framework is the same when a lot more subjects are recruited.
5.3. Quantification of Longitudinal Breast Cancer Changes

5.3.1. Introduction

According to American Cancer Society, breast cancer is the second most frequent cancer after skin cancer, and the second leading killer after lung cancer in US women (American Cancer Society, 2012). Patients with confirmed breast cancer often receive series of chemotherapies that span weeks to months. During such treatment, patients are often followed with periodic breast MRI scans to track cancer change and assess treatment effects. Registering these longitudinal breast MR images can help quantify the volumetric cancer changes. This section, therefore, applies the previously validated DRAMMS into quantifying the longitudinal breast cancer changes.

5.3.2. Material and Methods

Fourteen patients have been recruited from iSPY clinical trial. Image details can be found in Section 4.6.2.1. At the end of the 3-4 months chemotherapy course, these patients are further divided into two groups — responders, who show the absence of any residual invasive cancer in the breast and absence of any metastatic cells in the regional lymph nodes at the end of 3-4 months chemotherapy course, and non-responders otherwise.

Each subject’s follow-up images are registered to her baseline images by DRAMMS. Jacobian determinants of the obtained deformation are computed to show voxel-wise volumetric change. Jacobian determinant at each voxel is a measure of volumetric change ratio from the source to the target images. It shows expansion of a voxel if its value is greater than 1, shrinkage if its value smaller than 1 and volume-preservation if its value is exactly 1. To focus on volumetric changes in breast cancer regions only, breast cancer regions are
segmented from the baseline images by user-interactive intensity thresholding followed by human expert’s post corrections.

5.3.3. Results

Figs. 63 and 64 show the quantifications of longitudinal breast cancer changes in a typical non-responder patient and a typical responder patient. DRAMMS registration has enabled quantification of tumor change at voxel level. Compared with human experts who often report the volume change from the entire tumor regions, DRAMMS-quantified voxel-wise changes show more detailed information about which tumor sub-regions respond earlier and faster to chemotherapy. This valuable information has the potential use to predict a patient’s future response to chemotherapy during the course of treatment, or even before the start of the therapy. An accurate prediction can lead to a more timely adjustment of treatments, and a more effective treatment planning.

5.3.4. Discussion

This section applies DRAMMS to quantifying longitudinal breast cancer changes in individual patients. Based on accurately registering follow-up images to the baseline image, the Jacobian determinant maps calculated from the deformations reveal volumetric change at voxel level. Future study will on extracting cancer change patterns that have predictive values between responders and non-responders.
Figure 63: Quantification of the voxel-wise longitudinal breast cancer change by the DRAMMS registration. Shown here is a typical non-responder patient. Jacobian determinants in the tumor regions are overlaid on the baseline image in the bottom row.
Figure 64: Quantification of the voxel-wise longitudinal breast cancer change by the DRAMMS registration. Shown here is a typical responder patient. Jacobian determinants in the tumor regions are overlaid on the baseline image in the bottom row.
5.4. Multi-Atlas-based Skull-stripping for Brain Images

5.4.1. Introduction

Brain extraction, or skull stripping, is a very important pre-processing step preceding almost all automated brain MR imaging applications. It consists of the removal of the skull and the extra-cerebral tissues, e.g. scalp and dura, in the brain MR images. Brain extraction is known to be a difficult task, as the boundaries between brain and non-brain tissues, especially those between the gray matter and the dura matter, might not be clear on MR images. Also it is more prone to requiring manual intervention as the errors in this step propagate to most subsequent analysis steps, e.g., registration to a common space, tissue segmentation, cortical thickness estimation, etc.

Several brain extraction methods have been proposed since the early years of MR imaging research. One popular approach is known as BET (Smith, 2002), which is based on a deformable surface model that evolves to the boundaries of the brain. BET enjoys the advantage of being simple, fast, and general. But errors are not uncommon, since its assumption of a clear intensity contrast and a clear boundary between the brain gray matter and the dura matter is not always satisfied. Recently, atlas-based skull-stripping approaches have been explored. An atlas image, for which the ground-truth brain mask is provided, is registered to the target image space. The same deformation is used to warp the brain mask from the atlas space to infer the brain mask in the target image. To account for the large variation between the atlas and the target images, multi-atlas approaches are developed (Eskildsen et al., 2012). The promise is that registration of multiple atlases into the same target image may provide complimentary information to correct errors in each single-atlas-based brain mask propagation.

This section applies DRAMMS into multi-atlas-based brain extraction, or brain skull strip-
ping. With the high accuracy and robustness in within- and across-dataset registration of raw brain images (as shown in Section 4.2), DRAMMS is expected to increase the accuracy of the multi-atlas-based skull stripping framework. To demonstrate this, DRAMMS-based multi-atlas skull stripping results are compared to a recent method named ROBEX (Robust Brain Extraction) (Iglesias et al., 2011). ROBEX is chosen because it has been compared favorably against 6 state-of-the-art skull-stripping methods in multiple public datasets (Iglesias et al., 2011). Our experiment will show a high accuracy of our DRAMMS-based multi-atlas skull-stripping method compared with ROBEX in the same public datasets.

5.4.2. Material and Methods

Datasets. Three public datasets are used. They are also used in validating skull-stripping methods in (e.g., Iglesias et al. (2011)). They present different levels of difficulties, as we will see in the results section. The first is the Internet Brain Segmentation Repository (IBSR) dataset (Tsang et al., 2008). The IBSR dataset consists of 20 T1-weighted scans from the Center for Morphometric Analysis at the Massachusetts General Hospital. The brain was manually delineated by trained investigators in all scans. The second is the OASIS dataset (Marcus et al., 2007), which consists of 55 females and 22 males, with 51.64 ± 24.67 years of age. The brain masks were first generated by an automated method based on registration to an atlas, and then proofread and corrected by human experts before release. The third dataset is the LONI-LPBA40 dataset. It consists of 40 T1-weighted scans (20 males, 20 females, age 29.20 ± 6.30 years) and their corresponding annotations of brain masks.

Methods. Fig. 65 outlines the whole pipeline in our method. We apply our method on all subjects in the IBSR, OASIS and LPBA40 datasets. From each dataset, 10 atlases are selected. Atlas selection is by k-means \((k = 10)\) clustering of all subjects in that dataset based on an intensity-difference-based similarity between every pair of images after affine registration. By clustering, the selected atlases are expected to expand the variations of
Figure 65: Outline of our skull-stripping pipeline. In the registration component (middle), binary brain masks registered to target space are shown in yellow; in the label fusion component (bottom) the fused brain mask is shown using a blue-red colormap, and the final binary brain mask is shown in yellow.
subjects in that dataset. The selected atlases are then fixed for each dataset. All images except the selected ones are processed using the 10 atlases from the same dataset. These images selected as atlases are processed using 9 other atlases in the same dataset, so that the target and atlas images are not the same.

DRAMMS is used to register all atlases onto the target image. The obtained deformations are used to warp the brain extraction masks from the atlas to the target image space. To combine all the warped brain masks into a single brain mask, a weighted majority voting strategy is used. The weight is determined as inverse proportional to the absolute value of logarithm Jacobian determinant at each voxel. The assumption is that, at each voxel, we trust and weigh brain mask label more from the atlas that needs less deformation to arrive at this voxel in the target image (i.e., by smaller absolute value of logarithm Jacobian determinant of the deformation at this level).

**Accuracy Metrics.** We used the Dice score, the average surface distance errors (DE), the 95-percentile surface distance errors (95prc DE), the sensitivity and the specificity values between the automatically-computed brain masks and the ground-truth brain masks provided in these datasets. The results in these metrics are reported next to the ROBEX results from the same datasets, which are collected from the ROBEX paper [Iglesias et al., 2011].

**Sensitivity Analysis.** In all the above experiments comparing with ROBEX, we have selected 10 best atlases from each dataset. However, the final skull-stripping accuracy may vary if the number of atlases changes. So in additional sensitivity analysis, we have tested using different numbers (2–15) of atlases. This will show the stability of our methods with the varying numbers of atlases.
Table 6: Comparison of our skull-stripping results to ROBEX results in the same datasets. Results of ROBEX are from its own paper (Iglesias et al., 2011). DE stands for distance error between calculated and the ground-truth brain mask boundaries.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Dice</th>
<th>Ave DE</th>
<th>95prc DE</th>
<th>Sens.</th>
<th>Spec.</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBSR</td>
<td>ROBEX</td>
<td>95.6 ± 0.8</td>
<td>13.30 ± 2.60</td>
<td>3.80 ± 0.70</td>
<td>99.2 ± 50.0</td>
<td>92.3 ± 1.9</td>
</tr>
<tr>
<td>n = 20</td>
<td>ours</td>
<td>97.7 ± 0.8</td>
<td>15.01 ± 10.72</td>
<td>2.81 ± 1.10</td>
<td>97.5 ± 0.9</td>
<td>99.8 ± 0.2</td>
</tr>
<tr>
<td>OASIS</td>
<td>ROBEX</td>
<td>95.5 ± 0.8</td>
<td>9.80 ± 1.70</td>
<td>4.40 ± 0.60</td>
<td>93.8 ± 2.1</td>
<td>97.4 ± 12.0</td>
</tr>
<tr>
<td>n = 77</td>
<td>ours</td>
<td>96.1 ± 1.0</td>
<td>7.72 ± 1.51</td>
<td>3.81 ± 0.77</td>
<td>947 ± 2.6</td>
<td>99.2 ± 0.7</td>
</tr>
<tr>
<td>LPBA40</td>
<td>ROBEX</td>
<td>96.6 ± 0.3</td>
<td>13.30 ± 2.50</td>
<td>3.10 ± 0.40</td>
<td>95.6 ± 9.0</td>
<td>97.7 ± 7.0</td>
</tr>
<tr>
<td>n = 40</td>
<td>ours</td>
<td>98.2 ± 0.4</td>
<td>9.01 ± 4.07</td>
<td>1.95 ± 0.34</td>
<td>98.2 ± 0.6</td>
<td>99.7 ± 0.2</td>
</tr>
</tbody>
</table>

5.4.3. Results

Visual results for skull-stripping in a typical raw brain image have already been shown in Fig. 65. As all atlas-to-target registrations are within-dataset registrations, DRAMMS has a 0.934 ± 0.030 dice overlap between the warped and the target brain masks (please see within-dataset validation of DRAMMS in the same datasets in Fig. 38). This can be also seen from the high agreement among all warped brain masks in the second row of Fig. 65.

Quantitatively, Table 6 shows comparison of our results to ROBEX results. Our method has a slight edge in the OASIS dataset and bigger advantage in the IBSR and the LONI-LPBA40 datasets. As Figs. 34 and 38 have already shown in Section 4.2.3, the OASIS dataset has less anatomical variability and less FOV differences among subjects, and hence lower levels of difficulty for registration and atlas-based skull-stripping than the other two datasets. In two other more difficult datasets (IBSR and LONI-LPBA40), the edge of our method over ROBEX is bigger, underlining the effect of DRAMMS registration in compensating large cross-subject variations and FOV differences.

In addition, we have tested the sensitivity of our pipeline with regard to the number of atlases used. Fig. 66 shows the dice overlap of the calculated and the ground-truth brain masks using different numbers of atlases in our pipeline. Interestingly, the trend is very similar in different datasets. The accuracy of skull-stripping increases as the number of
atlases increases, but remains plateau after the number of atlases has reached 8 or 9. So 9 atlases can achieve good balance between accuracy and computational cost for all datasets used in this section.

5.4.4. Discussion

This section applies DRAMMS registration to the multi-atlas-based brain extraction (skull-stripping) framework. Results in multiple public datasets have shown consistently higher accuracy of our framework than the state-of-the-art brain extraction tool ROBEX (Iglesias et al., 2011). The accuracy mainly comes from multi-atlas settings and the accuracy and robustness in DRAMMS registration of raw brain images, which is one of the core parts in a multi-atlas-based segmentation framework.

Despite the high accuracy, there is still room for improvement in our pipeline.

Atlas selection is one component for future improvement. Right now a set of 10 atlases are fixed within each dataset. We can gain a higher accuracy by selecting target-specific atlases. The side effect is the increased computation and the requirement for ground-truth brain masks for far more subjects that could be potentially selected as atlases in the dataset. On the contrary, we can gain generality by fixing a number of atlases not for each dataset, but for all datasets. This makes a universal multi-atlas framework, saves computational time, and especially saves the need for labeling ground-truth brain masks for atlases in each dataset. However, it poses much higher requirement for the accuracy and robustness in the cross-dataset registration. Given DRAMMS’ accuracy and robustness in across-dataset registration of raw brain images (as demonstrated in Fig. 40 of Section 4.2.3), future work can be directed to such a general and practical framework.

Label fusion is another component that needs improvement. More sophisticated label fusion methods have been developed in the past decade, including the STAPLE (Warfield et al.,
Figure 66: Dice scores for different numbers of templates (atlases).
2004) tool and its variations (e.g., Commowick et al. 2012), and other advanced learning-based label fusion tools (e.g., Wang et al. 2011).

In a broad view, DRAMMS-equipped multi-atlas segmentation framework can be applied to segmentation of other anatomical structures, as long as DRAMMS is able to handle the atlas-to-target registration. Examples include multi-atlas segmentation of the prostate from the prostate MRI, left or right ventricles from cardiac images, and (sub)-cortical structures (e.g., hippocampus) from brain images.
5.5. Multi-Atlas-based Segmentation of the Prostate

5.5.1. Introduction

In MICCAI 2012 Prostate Segmentation Challenge, participants are provided with 50 training prostate MR images, each having expert-defined prostate segmentation. The task is to segment prostate in 30 testing prostate MR images. In such a setting, we present a multi-atlas-based segmentation pipeline. The central idea is to transfer those expert-segmentations in training images (i.e., atlases) onto target image through image registration, and then fuse the transferred segmentations to derive an ultimate prostate segmentation in the target image. Previous studies (e.g., [Klein (2008); Martin et al. (2008); Gubern-Merida and Marti (2009); Dowling et al. (2009)]) have reported plausible results using multi-atlas strategy to segment prostate, with average dice of 0.75–0.80 compared to expert-defined ground-truth.

In the dataset used in MICCAI 2012 Prostate Segmentation Challenge, several factors pose large difficulties to image registration, which is the fundamental component in multi-atlas segmentation framework. Those factors include: large variability of the MR images not only in terms of image intensity characteristics (e.g. scanner variability, inhomogeneities, etc.) but also in structures. Images may often have different field of views (FOVs) and are obtained from different imaging centers (see Fig. [69] for example). Consequently, the registration, affine or deformable, may often times fail. Aside from anatomical variabilities, this dataset contain some unidentified subjects with enlarged prostate or even prostate cancers, causing pathological variabilities.

To overcome these limitations, the proposed method is based on a zooming process: atlas-to-target registration to land an initial segmentation of prostate; then "zoom in", focusing only on the vicinity of prostate and re-do atlas-to-target registration to increase accuracy.
in registration and hence segmentation.

The key to guarantee the success of this zooming process is to obtain reasonably good initial segmentation of the prostate before zooming in to its vicinity. To this end, a registration method that is robust against the image variation is used, followed by iterative label fusion and atlas selection (it is not uncommon to expect certain number of registrations to fail, therefore we need to have an automatic mechanism to remove those failed cases from label fusion). Our experiments in training dataset will show that this strategy shall yield an initial segmentation of prostate that is of average 0.81 dice overlap with expert-segmentation (see Fig. 68 "phase 1" results). Obtaining this reasonably high accuracy is important, as we can now zoom in to the dilated version of the initial segmentation mask and expect the whole prostate being included in this dilated mask. The remaining part is straightforward – with much less interference from those structures far away from the prostate, we can focus on registering prostate vicinity well and finally obtained a refined final segmentation of the prostate. This zooming process has lifted the average dice to 0.84 in our experiment.

In the following, we will describe details of our pipeline in Section 5.5.2 with some demonstrations in the training dataset in Section 5.5.3 Section 5.5.4 discusses and concludes this section.

5.5.2. Methods

We will detail the proposed zooming process in this section. The whole pipeline is depicted in Fig. 67. In this figure, light pink box contains the first phase, where whole-image atlases are registered to whole-image target to land a tentative segmentation of the prostate (e.g., the orange region in the rightmost figure within pink box). Then, in the second phase, we zoom into prostate and its vicinity, as shown in the gray box. Here we refine all registrations and obtain the final segmentation of the prostate (e.g., the red region in the rightmost figure within gray box). The following two sub-sections will describe each of those two phases.
Figure 67: The proposed pipeline. We rely on the whole atlases for initial segmentation of the prostate (phase 1, pink box), and then zoom in to the vicinity of prostate to obtain the final multi-atlas-based segmentation (phase 2, gray box).
### 5.5.2.1. Initial Multi-Atlas Segmentation: Find Prostate Location

In this first phase, we will use whole image registration from multiples atlases to target to obtain an initial segmentation of the prostate. Due to large variations in imaging protocols, FOVs, structures, anatomies and even pathology conditions among different subjects, registration from an atlas to the target is a very different task (see Fig. 69 for example). To ease this difficulty, we will describe below two measures: a) to use a robust registration algorithm; b) to automatically select atlases and potentially remove atlases whose registrations to target image have failed.

For registration algorithm, a recently-developed non-rigid registration algorithm is used for warping atlas images to the target. This algorithm, termed DRAMMS registration [Ou et al. (2011)], is based on an attribute-based similarity metric. That is, it finds voxel correspondences by rich set of geometric texture at each voxel, other than by image intensity alone. The high dimensional multi-scale and multi-orientation Gabor textures have rendered each imaging voxel more distinctive and therefore better identifiable during search for correspondence. Furthermore, when registering an atlas to the target image, this algorithm relies more on the regions that can establish a more reliable matching compared to other regions. Such an approach is particularly well suited to the registration of prostate images, where the two images may have significant differences, or even missing correspondence (i.e., some structures present in one image but not the other).

For atlas selection and removal, we shall keep those atlases that are similar to target image and on the other hand, remove those atlases whose registrations to target image have failed. We automate this process in an iterative fashion. At the very beginning, we used all atlases for label fusion (using a population label fusion method STAPLE [Warfield et al. (2004)]). After having obtained a tentative prostate segmentation, we measure atlas-to-target similarity by a) mutual information (MI) within the tentative segmented region, b) correlation coefficient (CC) within the tentative segmented region and c) dice overlap
between warped prostate region and the tentative prostate segmentation. Those atlases having less than 90% of the maximum MI, less than 90% of the maximum CC and less than 50% of the maximum dice overlap among all atlases are considered having failed registration with target image and thus removed. The remaining atlases are weighted by their $MI \times CC$ values in a globally weighted majority voting label fusion. The assumption here is to trust more on those atlases having higher similarity with the target image after registration. This way, we get an updated prostate segmentation. We iterate between label fusion and atlas selection/ranking until convergence. The convergence criterion is met if the prostate segmentations in two consecutive iterations become relatively stable (> 90% dice), or otherwise repeat the above process till convergence. After convergence, the prostate segmentation mask in this phase 1 will be used for initializing phase 2 (focusing on prostate vicinity) in the next subsection. All the above mentioned parameters are optimized by leave-one-out experiments in training dataset.

5.5.2.2. Final Multi-Atlas Segmentation: Focus on Prostate Vicinity

The tentative prostate segmentation by the above phase 1 has an average 0.81 dice overlap with expert-define ground truth. This shows that phase 1 has successfully located majority part of the prostate in the target image.

With this reasonably good localization of the prostate in target image, we can now zoom in and focus on prostate vicinity in both atlas and target images. We first isotropically dilate the prostate mask obtained from phase to 1.5 times its volume, to make sure the dilated mask can cover the whole prostate (together with other surrounding structures). This way we have zoomed into prostate vicinity in the target image. To also zoom into prostate vicinity in atlases images, we proportionally dilate the ground-truth prostate masks in atlases. Readers can follow the arrows pointing from pink box to gray box in Fig. 67 to visualize this zooming process for extracting the prostate vicinities in both atlases and target images. It should not be difficult to observe the immediate benefits of this zooming process.
-- a large part of registration difficulties caused by different FOVs, different structures, and to some extent image inhomogeneity have been removed. Now we only need to register prostate vicinity in atlas and target images, which is a much easier problem than registering the whole images. Here we have also used DRAMMS registration software [Ou et al. (2011)]. The resultant warped prostate masks are fused by the same strategy we have used in phase 1 (majority voting with iterative atlas ranking and selection).

5.5.3. Results

To demonstrate the effect of atlas ranking/selection and the effect of zooming process, we compared accuracies obtained from 4 different multi-atlas segmentation methods: MV -- majority voting based on all warped atlases, no zooming process; STAPLE -- based on all warped atlases, no zooming process; the proposed phase 1 -- the proposed iterative weighted MV label fusion together with atlas ranking and selection, but no zooming process; the proposed phase 1+2 -- the entire proposed pipeline, including both phase 1 (initial prostate localization) and phase 2 (zooming process with focus on prostate vicinity).

Fig. 69 visualizes segmentation by our proposed pipeline (green contours) as well as ground-truth segmentation (red contours), for 6 typical training subjects. These 6 subjects are randomly chosen from different imaging centers. For each subject, we used the remaining 49 subjects in training as atlases. During atlas selection, usually 10-15 atlases will be automatically kept. Fig. 68 quantifies segmentation accuracy in those cases, as measured by dice overlap with ground-truth segmentation.

We have several interesting findings from Fig. 68. First of all, atlas ranking and selection improve segmentation accuracy. This is expected, as removing those atlases that fail to register to target image should reduce confusions in label fusion. Also, the proposed zooming process does offer additional improvement in segmentation accuracy, accompanied by reduced standard error, showing the advantage of focusing on prostate vicinity.
Figure 68: Segmentation accuracy in 6 typical training subjects. Here accuracy is measured by dice overlap with expert-defined ground truth segmentation.

5.5.4. Discussion

This section proposed a fully automatic pipeline for segmenting prostate MR images. We made two contributions in this study. The first is using robust image registration and especially atlas selection in multi-atlas-based segmentation framework. Due to large variability in prostate images across subjects, failure in atlas-to-target registration is not uncommon. Therefore, using robust image registration method to reduce the number of failures and developing criterion for removing failed atlases become key to accurate segmentation.

Our second contribution is the development of a zooming process and its proof of concept. The idea is to focus on registration of the prostate vicinity, in the hope of improving registration accuracy in prostate regions and hence the accuracy in segmentation. By doing so, we largely reduce the negative impact from those highly variable structures far away from the prostate. As a result, we observed significant improvement in segmentation accuracy.

We have applied this automatic pipeline to all 30 testing images in the MICCAI 2012 Prostate Segmentation Challenge. Results will be disclosed by the organizers of this challenge.
Figure 69: Visualization of segmentation results in 6 randomly chosen subjects from training set. Red contours are ground truth by expert segmentation. Green contours are generated by our proposed pipeline.
5.6. Conclusion

This chapter demonstrates the wide application of the extensively validated DRAMMS registration method. In particular, DRAMMS has been applied to the five representative topics from three major branches of studies where image registration plays a central role.

Constructions of statistical lesion atlases and tumor atlases are representative topics in population studies. In population studies, registering images from a population of subjects into a common template image space enables joint analysis of all subjects. This can lead to discovery of the common, but often subtle, patterns of certain abnormalities in a specific population. It can also lead to the discovery of certain image biomarkers that can differentiate between sub-populations (e.g., normal v.s. diseased). In addition, it can also lead to the discovery of where certain abnormalities are most likely to occur in specific populations (e.g., where prostate cancer most likely to occur in Asian population). Such discoveries can provide prior knowledge for the targeted abnormality detection (e.g., the targeted biopsy strategy [Ou et al. 2009b]). DRAMMS is readily applicable in the population studies, thanks to its demonstrated accuracy and robustness with regard to cross-subject anatomical variations, and even to partial missing correspondences that are not rare in pathologic population.

Quantification of the breast cancer change is a representative topic in longitudinal studies. In longitudinal studies, registering images of the same subject from different time points captures how images deform over time. It can lead to the tracking of structural and functional growth (such as in pediatric imaging or in mouse brain development experiments). It can also lead to the monitoring of treatment effects in longitudinal images of various cancer types (brain tumor, breast cancer, prostate cancer, etc). DRAMMS is readily applicable in many of the longitudinal studies, because of its demonstrated registration accuracy and its ability to deal with large structural changes over time.
The multi-atlas-based segmentation of the brain and the prostate are representative topics in atlas-based segmentations. Atlases contain the prior knowledge of anatomical structures, which can be propagated to the target image by the accurate registration. This is based on the assumption that atlases are anatomically similar to target images. One fundamental challenge in atlas-based segmentation is the violation of such an assumption. In such cases, registration between the largely variant atlas and target images may fail, hindering the propagation of structural segmentations. DRAMMS is applicable in many of multi-atlas segmentation tasks, because of its demonstrated accuracy and robustness to within-dataset, and especially across-dataset variations.

Despite the applications mentioned above, it is important to note, at the end of this chapter, some situations where DRAMMS cannot directly apply. In general, those situations often present fairly challenging registration problems caused by considerable anatomical variations. For instance, in population study, it is hard for DRAMMS to register a population of raw cardiac images into a common cardiac template space. This is mainly because large different FOVs and loss of anatomical correspondences in non-cardiac structures that often dominant the entire images. In longitudinal studies, direct registration of images from time points too far away from each other may also pose challenges to DRAMMS and almost all other image registration methods. Registering mouse brain images from early days (childhood) to senior age is one such example where many registration methods including DRAMMS may fail to capture change. In such situations, registrations along a path of time points with gradual anatomical changes can help alleviate the problem.

In conclusion, this chapter presents applications of DRAMMS registration in representative population studies, longitudinal studies and (multi-)atlas-based anatomical segmentation tasks. In the next chapter, we will attempt to further improve the speed and accuracy of the DRAMMS framework.
CHAPTER 6

Towards Improving DRAMMS Speed and Accuracy

This chapter makes two attempts towards further improving the speed and the accuracy of the DRAMMS framework.

The first attempt, in Section 6.1, tries to reduce the computational cost and improve the speed of DRAMMS. Instead of establishing correspondences at all voxels in the DRAMMS framework, a discreteDRAMMS framework is proposed to extract landmark pairs at a number of automatically-computed voxel locations. The landmark pairs preserve the accuracy of DRAMMS, and largely reduce the computational cost. They have the potential to guide other intensity-based registration methods, which are computationally more efficient than DRAMMS.

Section 6.2 is the second attempt. It tests the idea of incorporating task-specific prior knowledge into DRAMMS to further improve the registration accuracy. Specifically, we test the effect of the automatic brain tissue segmentation, which is a form of prior knowledge about brain tissue membership, to the accuracy of registering normal brain images.
6.1. discreteDRAMMS: Landmark Detection and Matching

6.1.1. Introduction

The high-dimensional attribute matching and mutual-saliency weighting components in DRAMMS help improve the accuracy of registration. However, they also increase computational cost compared with intensity-based registration methods. Our long-term goal is to combine the accuracy in DRAMMS and the computational efficiency in intensity-based methods.

Towards this long-term goal, a hybrid approach is proposed. It first converts DRAMMS into discreteDRAMMS — establishing correspondences at a set of automatically-located voxels other than all voxels in DRAMMS. The immediate advantage is the reduced computational cost. Then, we can use the correspondences from discreteDRAMMS, which preserves the accuracy of DRAMMS, to guide intensity-based registration methods, which is computationally efficient.

This idea is similar to some recently-developed hybrid registration methods that also combine landmark information with intensity-based voxel-wise methods (e.g., (Hellier et al., 2003; Sotiras et al., 2010; Johnson and Christensen, 2002)). They show that the intensity-based registration methods obtain higher registration accuracies when guided with the landmark-induced geometric information. The main difference between the proposed work and other works are in two aspects. One is how to automatically extract landmark correspondences (the proposed discreteDRAMMS herein). The other is how to combine landmark-induced geometric information into intensity-based methods. Since the second aspect has already been studied in our collaborative work (Sotiras et al., 2010), this section will fo-
cus on the first aspect — discreteDRAMMS as a new approach to automatically extract landmark correspondences.

Establishing landmark correspondences usually involves the detection of representative landmark locations and the matching of them between images according to some similarity measure. Landmark detection and matching, however, are non-trivial tasks that often face three challenges described below.

The first and foremost challenge is the outlier problem. An outlier is a landmark that cannot find its true counterpart in the landmark set extracted from the other image. Outliers may arise due to failures in detection and matching steps, or due to the lack of correspondences in the presence of pathologies or simply cross-subject variations. In the literature, the Soft-Assign algorithm (Chui and Rangarajan 2003), developed after its precedent Iterative Closest Point (ICP) algorithm (Besl and McKay 1992), introduced fuzzy correspondences and attempted to handle outliers by an explicit exclusion mechanism. However, it requires iterations of computationally intensive transformation and is relatively sensitive to initial conditions.

The second challenge is the self-crossing problem in the implied dense deformation. This often arises when spatially nearby landmarks find their correspondences in reverse directions. The fundamental problem is in finding correspondences for nearby landmarks independently without considering their joint influence in the regions they represent. One solution is to strictly constrain diffeomorphic correspondences during iterative transformation (Joshi and Miller 2000), yet despite the computational load, the strict constraint may or may not hold under inter-subject variability.

The third challenge is non-uniform spatial distribution of landmarks. As a result, regions having less or even no landmarks tend to have degraded accuracy compared to regions having more landmarks. Although strictly uniform distribution of landmarks is often impractical due to the nature of objects in the images, it is still preferred to have spatially dispersed
landmarks to offer distributed registration accuracy in the image space.

The discreteDRAMMS approach is designed in this section to meet these aforementioned challenges. To alleviate outlier problem, the proposed framework consists of Gabor texture-based landmark detection and an initial matching by forward-backward one-to-one uniqueness constraints. It is based on the effective mutual-saliency measure in the DRAMMS framework. Following this, a spatial clustering component is specifically designed in the refined matching step to meet all three desirable properties: the further removal of outliers if any, the deformation smoothness and the landmark dispersion. In addition, our framework is completely symmetric regardless of which image of the two is chosen as the template (theoretically guarantees the same results when permuting them). We demonstrate these properties in simulated data, and show visual results in real data involving tumor-induced missing correspondences.

6.1.2. Method

The proposed discreteDRAMMS approach has three steps: landmark detection, the initial matching by uniqueness constraints, and the refined matching by spatial clustering. Fig. 70 shows representative results following each step. We elaborate the whole framework below.
Step 1: Landmark Detection. Two 3D intensity images $I_1 : \Omega_1 \subset \mathbb{R}^3 \mapsto \mathbb{R}$ and $I_2 : \Omega_2 \subset \mathbb{R}^3 \mapsto \mathbb{R}$ are convoluted with a set of 3D multi-scale and multi-orientation Gabor filters. Voxels having the highest magnitude in the Gabor response are detected as landmarks. Here a threshold for magnitude is adaptively set at 85-percentile of the histogram of the Gabor responses, so that around 15% of all voxels will be included, as shown in Fig. 70(a). Other thresholds (5% and 10%) were also tested but made no big difference in final results. The detected landmarks usually correspond to edge/boundary voxels or regional centers at various scales and orientations. Other landmark detection methods, such as Canny, LoG, SIFT, can also be used here; we choose Gabor because of its favorable distinctiveness in multi-scale and multi-orientation fashion and its demonstrated accuracy in image registration (Liu et al., 2002). The fact that Gabor filters are not affine invariant is not a big concern, because an initial affine registration has already removed global affine differences. Other than this initial affine registration, our framework hereafter assumes neither image segmentation nor transformation.

Without loss of generality, we denote those $N_1$ and $N_2$ (usually $N_1 \neq N_2$) landmarks detected from two images as two sets $L^{(1)} = \{l_i^{(1)}\}_{i=1}^{N_1}$ and $L^{(2)} = \{l_j^{(2)}\}_{j=1}^{N_2}$, where subscript indexes landmark and superscript indexes image space.

Step 2: Initial Matching by Uniqueness Constraints. Some detected landmarks cannot offer good correspondences. To remove them, we design uniqueness constraints in this initial matching step.

Before describing uniqueness constraints, we first need to define "matching degree (MD)" to quantify the likelihood of two landmarks $l_i^{(1)} \in I_1$ and $l_i^{(2)} \in I_2$ to be true correspondences. We follow the definition in (Ou et al., 2011),

$$\text{MD}(l_i^{(1)}, l_j^{(2)}) = \text{Similarity}(l_i^{(1)}, l_j^{(2)}) \times \text{MutualSaliency}(l_i^{(1)}, l_j^{(2)})$$ (6.1)

where the similarity is based on their Gabor attributes, and mutual-saliency quantifies
their matching reliability, or confidence in their matching. For more details on similarity and mutual-saliency definition, please refer to Sections 3.2.2.2 and 3.2.3 in this dissertation.

Having defined matching degree (MD), we are ready to define "uniqueness constraints" as follows: out of all \( N_1 \times N_2 \) possible pairs, we keep those pairs \((l^{(1)}_m, l^{(2)}_n)\) for all \( m \) and \( n \), such that

\[
m = \arg \max_i MD(l^{(1)}_i, l^{(2)}_m) \quad (6.2)
\]

and

\[
n = \arg \max_j MD(l^{(1)}_m, l^{(2)}_j) \quad (6.3)
\]

where Eqn. 6.2 ensures that the source point \( l^{(1)}_m \) is the best match for the target point \( l^{(2)}_n \) out of all source candidates \( i \)'s – what we call "forward uniqueness"; and reversely, Eqn. 6.3 ensures that the target point \( l^{(2)}_n \) is the best match for the source point \( l^{(1)}_m \) out of all target candidates \( j \)'s – what we call "backward uniqueness".

An immediate advantage is the removal of most outlier pairs (typical results shown in Fig. 70(b)). This is because, an outlier landmark may occasionally have high similarity with another landmark, but it will be less likely to have high matching reliability in the neighborhood, and far less likely to have the forward-backward one-to-one uniqueness as required in equations 6.2 and 6.3.

**Step 3: Refined Matching by Spatial Clustering.** Following the initial matching, this refined matching step aims at three desirable properties: further removal of outliers if any, spatial smoothness and spatial dispersion of landmarks.

These three properties are approached by three terms in a clustering formulation, where we cluster all candidate pairs left from initial matching, and use cluster exemplars as final landmark correspondences.

Note that, the clustering is on landmark pairs after initial matching, not landmark individuals. Each pair contains source and target landmark locations, a matching degree (MD)
and an implied displacement – all this quantitative information will be used in the spatial clustering.

Mathematically, suppose we have $H$ candidate pairs kept from initial matching, denoted as a set $\mathcal{P} = \{ p_h | p_h = (l_h^{(1)}, l_h^{(2)}), h = 1, \ldots, H \}$. Our objective is to choose a number of $K$ cluster exemplars $\{ c_k \}_{k=1}^{K}$ to satisfy those three desirable properties. Moreover, this number $K$ should be automatically determined by the clustering algorithm itself. Analogous to Affinity Propagation (AP) [Frey and Dueck 2007] and Linear Programming Stability-based (LP-stability) [Komodakis et al. 2008] clustering algorithms, we define an overall energy function below:

$$\min_{K, \{c_k\}_{k=1}^{K}} \left( \sum_{p \in \mathcal{P}} \min_k d^A(p, c_k) + \sum_{r=1}^{K} \left( d^B(c_r, c_r) + \sum_{s=1}^{K} d^\Gamma(c_r, c_s) \right) \right) \quad (6.4)$$

The first term in Eqn. 6.4, $d^A(p, c_k)$, encourages spatial smoothness,

$$d^A(p, c_k) = \frac{1}{2} \| (l_p^{(1)} - l_{c_k}^{(1)}) \| + \frac{1}{2} \| (l_p^{(2)} - l_{c_k}^{(2)}) \| + \| (l_p^{(2)} - l_{c_k}^{(1)}) - (l_{c_k}^{(2)} - l_{c_k}^{(1)}) \| \quad (6.5)$$

where bold letter $l$ represents spatial coordinates of landmark $l$, subscript indexes landmark and superscript indexes image space ((1) for source and (2) for target). In particular, it requires a) the source points in pair $p$ and its exemplar $c_k$ having spatially close locations $l_p^{(1)}$ and $l_{c_k}^{(1)}$; b) the target points in pair $p$ and its exemplar $c_k$ also having spatially close locations $l_p^{(2)}$ and $l_{c_k}^{(2)}$; and c) nearby pairs $p$ and $c_k$ having smooth displacements $(l_p^{(2)} - l_p^{(1)})$ and $(l_{c_k}^{(2)} - l_{c_k}^{(1)})$.

The second term in Eqn. 6.4, $d^B(c_r, c_r)$, further removes outliers if any. It chooses cluster exemplars $c_r$ as ones having high matching degree (MD) defined in Eqn. 6.1, therefore it is defined as

$$d^B(c_r, c_r) = \frac{1}{\text{MD}(l_{c_r}^{(1)}, l_{c_r}^{(2)})} \quad (6.6)$$
The third term in Eqn. 6.4, $d^G(c_r, c_s)$, encourages spatial dispersion of selected landmark pairs. Specifically, it requires any two selected exemplars $c_r$ and $c_s$ to be spatially distant in source and target points, and is therefore defined as:

$$d^G(c_r, c_s) = \left( \|l^{(1)}_{c_r} - l^{(1)}_{c_s}\| + \|l^{(2)}_{c_r} - l^{(2)}_{c_s}\| \right)^{-1}. \tag{6.7}$$

Note that the number of clusters (hence finally selected landmark correspondences) $K$ is also optimized rather than pre-set by users. We have experimented with both Affinity Propagation (Frey and Dueck, 2007) and LP-Stability clustering methods (Komodakis et al., 2008), which we finally used, and obtained almost the same results.

The formulation of all three steps in our framework is completely symmetric regarding the choice of source and target between two images, so the same results are guaranteed when permuting them.

### 6.1.3. Results

We present both quantitative results in simulated data and visual results in real images. Overall, our framework runs in about 10 minutes for typical 3D brain images or about 5 seconds for 2D images on 3.4GHz CPU, and usually identifies 100–300 pairs of landmark correspondences in 3D or 10–25 pairs in 2D.

#### 6.1.3.1. Demonstration of Three Properties in Simulated Data

Eight normal brain images ($256 \times 256 \times 181$ voxels, $0.9375 \times 0.9375 \times 1.0 \text{mm}^3/\text{voxel}$) were used to generate eight deformed images, by simulating a deformation field in each image space using a wavelet PCA-based deformation simulator (Xue et al., 2006). The simulated displacement is on average 6-9 mm per voxel in each simulated deformation field. To further
test robustness of the proposed framework, 5% and 10% Gaussian noises were also added into the simulated images (percentage is in reference to the greatest intensity in the image). Landmark correspondences are extracted between each pair of the original and simulated images.

Validation of Property 1: Accuracy and Robustness to Outliers. We measure the errors of landmark-implied displacements with regard to the ground truth deformation at the extracted landmark locations. We compare the errors before and after clustering step, which is a major feature in our framework, and further compare with errors at the same final locations by two dense registration methods Demons (Vercauteren et al., 2009) and DROP (Glocker et al., 2008). Comparison to Demons and DROP is because of their demonstrated accuracy, and because it can reflect to which extent the landmark-implied geometric information can ameliorate dense registration, which is our ultimate goal.

Comparison results in Fig. 71 prompt three observations. First, our framework (purple bars in the figure) has obtained lower errors, which manifests its potential in guiding dense registration. Second, this trend is more clear in more difficult cases (middle and bottom sub-figures), showing robustness of our methods amidst image noise and inhomogeneities. Third, comparison before and after clustering (green and purple bars) underlines the importance of the clustering step in further removing outliers (reduced standard deviation of errors) and increasing accuracy (reduced mean error).

Validation of Property 2: Spatial Smoothness in the Implied Field. Landmark correspondences before and after clustering are used to interpolate a dense deformation by thin-plate-spline (TPS) strategy (Rohr et al., 2001). Table 7 shows smoothness in the deformation field: a) the considerable reduction of negativity in minimum Jacobian determinants implies reduced self-crossing in deformation field; b) smaller deviation of min and max Jacobian determinants from value 1 (i.e., no deformation at all) reflects less bending and thus increased smoothness.
Figure 71: Displacement errors compared to ground-truth deformation. Top: between original and simulated images; Middle: between original image and simulated image with 5% noise; Bottom: between original image and simulated image with 10% noise.

Table 7: Smoothness measured by min and max Jacobian determinants in the dense deformation fields interpolated by landmark correspondences before and after clustering.

<table>
<thead>
<tr>
<th></th>
<th>min Det(Jacobian)</th>
<th>max Det(Jacobian)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BeforeClustering</td>
<td>AfterClustering</td>
</tr>
<tr>
<td>no noise</td>
<td>-2.67±1.57</td>
<td>0.05±0.53</td>
</tr>
<tr>
<td>5% noise</td>
<td>-1.74±1.30</td>
<td>0.16±0.89</td>
</tr>
<tr>
<td>10% noise</td>
<td>-2.46±1.48</td>
<td>0.06±1.02</td>
</tr>
</tbody>
</table>
Table 8: Spatial dispersion of landmark correspondences before and after clustering, measured by the trace of the covariance matrix of all landmark locations.

<table>
<thead>
<tr>
<th></th>
<th>BeforeClustering</th>
<th>AfterClustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>no noise</td>
<td>736.61±141.18</td>
<td>807.99±109.61</td>
</tr>
<tr>
<td>5% noise</td>
<td>736.74±132.89</td>
<td>814.37±103.21</td>
</tr>
<tr>
<td>10% noise</td>
<td>809.45±145.09</td>
<td>906.47±125.97</td>
</tr>
</tbody>
</table>

**Validation of Property 3: Spatial Dispersion.** In Spatial Descriptive Statistics, spatial dispersion is often measured by the trace of covariance matrix of the data – higher trace indicates higher total variance and hence higher degree of dispersion. By adopting this concept, Table 8 shows an increased degree of landmark dispersion as a result of spatial clustering. This highlights the capability of our framework to provide more distributed geometric guidance to dense registration.

### 6.1.3.2. Results in Real Data

The previous Fig. 70 has already shown results in real brain MR images. In addition, our framework is also applied to real images containing tumor-induced loss of correspondences, which poses challenges to most registration methods.

A typical set of results in Fig. 72 shows the capability of our framework to provide visually accurate and spatially distributed landmark correspondences in regions where correspondences can be found while circumventing missing correspondences. This shows its robustness and its potential to offer geometric guidance even in normal-to-pathology registration tasks.

### 6.1.4. Conclusion

This section presents a framework for landmark detection and matching between two images. Compared with other methods, our framework is free of segmentation or transformation,
Figure 72: Landmark correspondences established in real images containing tumor-induced missing correspondences. (a) results of landmark detection; (b) results of initial matching; (c) final results after spatial clustering. Blue arrow points out tumor. Red crosses label extracted correspondences. Cross size indicates matching degree, or reliability.

and is completely symmetric with regard to source and target image spaces. Furthermore, it has three features in the extracted landmark correspondences: accuracy and robustness to outliers, spatial smoothness in implied dense deformation, and spatial dispersion of landmarks to offer distributed geometric guidance to dense registration. We have quantitatively demonstrated these features in simulated brain MR data, and visualized its robustness in real data containing tumor-induced missing correspondences.

Future work includes testing our framework in larger datasets involving more challenging deformations. The endpoint is a coupling of our landmark-implied geometric information with dense registration. It will heritage registration accuracy and robustness in DRAMMS and computational speed in intensity-based registration methods. This will expand our ability in brain function study, in computer-aided pathology diagnosis and in image-based treatment planning.
6.2. Tissue Segmentation Improves Brain Registration Accuracy?

6.2.1. Introduction

DRAMMS was designed as a general-purpose registration method. Therefore it does not require any prior knowledge. However, DRAMMS can be extended to incorporate prior knowledge in specific registration tasks. That is also one reason and advantage of designing DRAMMS as a general-purpose method in the first place. Prior knowledge can be expert-defined landmarks, can be expert-defined ROIs, can be the range or regularization on deformations, and so on. As a starting point, this section focuses on the classic brain image registration problem, and focuses on the prior knowledge from automatic tissue segmentation in brain images. In the following, we study whether automatic tissue segmentation will further improve registration accuracy in cross-subject brain image registration.

The past two decades have witnessed extensive research on image registration. Besides classifying them into landmark-based and voxel-wise methods, we can also classify registration methods into segmentation-free and segmentation-dependent. Segmentation-free methods only use gray-scale intensity information for registration (Avants et al., 2008; Rueckert et al., 1999; Vercauteren et al., 2009). Segmentation-dependent methods use the segmentation to either initialize registration (Heckemann et al., 2010; Ashburner, 2007; Shen and Davatzikos, 2002), or interleave with registration in a joint framework (Pohl et al., 2006; Wang et al., 2006; Wyatt and Nobel, 2003). Of those methods, a natural question is, whether segmentation improves registration accuracy, and if so, to which extent.

Studies (Yassa and Stark, 2009; Pohl et al., 2006) showed that structural segmentation does improve registration accuracy. Here structure segmentation refers to segmentations of
localized sub-cortical brain structures, such as hippocampus, perirhinal cortex, entorhinal cortex or alike. However, studies are lacking as whether tissue segmentation also improves registration accuracy. Here, tissue segmentation often refers to segmentation of brain images into three tissue types: white matter (WM), gray matter (GM), and cerebrospinal (CSF). Compared with structural segmentation, tissue segmentation is a more extensively studied topic with more established tools (e.g., (Zhang et al., 2001; Ashburner and Friston, 2005)) to obtain greater accuracy and robustness. Therefore, investigating whether tissue segmentation improves registration is a more general problem with likely more applications.

Evaluating the effect of automatic tissue segmentation on registration is, however, not a trivial problem. Direct comparison between those methods using (eg, (Ashburner, 2007; Shen and Davatzikos, 2002)) and those methods not using (eg, (Avants et al., 2008; Rueckert et al., 1999; Vercauteren et al., 2009)) tissue segmentation is often unfair, since they also differ in other compounding factors, such as optimization strategy, transformation model and regularization levels. A fair evaluation should be to find a "testbed" registration method, fix all those compounding factors, and then compare the deformation accuracies between feeding it with gray-scale intensity images and feeding it with tissue segmentation images. Under this strategy, Heckemann et al (Heckemann et al., 2010) in a recent study reported improved success rate and improved accuracy of registration as a result of tissue segmentation. But this is specifically for B-spline method (Rueckert et al., 1999), and significance of improvement seems not clear.

To make a more general observation, this section conducts a larger experimental study. 28,800 registrations have been run by 8 popular registration methods in 2 large-scale public datasets. Results help to show whether tissue segmentation helps different registration methods in different datasets, and to which extent. It will provide references to the question that whether automatic tissue segmentation will further improve DRAMMS registration accuracy in cross-subject normal brain image registrations.
6.2.2. Material and Method

6.2.2.1. Overview of Experiment Design

Fig. 73 outlines our study. It can be interpreted as follows. Without loss of generality, our purpose is to examine, for a given "testbed" registration method $M$, whether tissue segmentation will improve registration from subject $i$ to subject $j$ ($i \neq j$). We approach this purpose by the following experiment. Suppose intensity images of those two subjects are $I_i$ and $I_j$. From those intensity images, we can compute automatic tissue segmentation images $S_i$ and $S_j$. Now, we have two ways to register from subject $i$ to subject $j$. On way is to use intensity images; suppose the result deformation is $f^M_{I_i \rightarrow I_j}$. The other way is to use tissue segmentation images, suppose the result deformation is $f^M_{S_i \rightarrow S_j}$. To this end, the problem of whether tissue segmentation improves registration accuracy can be boiled down to the question whether deformation $f^M_{S_i \rightarrow S_j}$ is more accurate than deformation $f^M_{I_i \rightarrow I_j}$. If deformation $f^M_{S_i \rightarrow S_j}$ is more accurate, then tissue segmentation improves registration accuracy, and vice versa. Here, accuracy of deformation $f^M_{S_i \rightarrow S_j}$ is indicated by whether it can warp expert-annotation $L_i$ of subject $i$ into agreement with expert-annotation $L_j$ of subject $j$. That is, accuracy of deformation $f^M_{S_i \rightarrow S_j}$ is indicated by Jaccard Overlap $JO(f^M_{S_i \rightarrow S_j} \circ L_i, L_j)$. In the same manner, accuracy of deformation $f^M_{I_i \rightarrow I_j}$ can be indicated and comparison of two deformations can be made.

In this framework, we need to describe 1) what datasets we have used, which can be found in Section 6.2.2.2 2) which registration methods we have recruited as testbeds, which can be found in Section 6.2.2.3 3) how to perform automatic tissue segmentation, which can be found in Section 6.2.2.4 and 4) how to measure registration accuracy, which can be found in Section 6.2.2.5.
Figure 73: Overview of the experiment design in this study. Please refer to Section 6.2.2.1 for a brief interpretation.
6.2.2.2. Datasets

Two public brain MRI datasets are used in this study. The first one is the NIREP dataset (Christensen et al., 2006) that was launched by researchers in the University of Iowa. The NIREP dataset contains T1-weighted brain MR images of 16 healthy subjects (image size $256 \times 300 \times 256$ and voxel dimension $1.0 \times 1.0 \times 1.0\, \text{mm}^3$). Each NIREP subject has expert annotation of 32 localized structures; they are distributed in the frontal, parietal, temporal and occipital lobes, cingulate gyrus, and insula. We refer to (Christensen et al., 2006) for a complete list of all those structures. The second dataset we used is the LPBA40 dataset (Shattuck et al., 2008) that was launched by researchers in UCLA. The LPBA40 dataset contains T1-weighted images of 40 healthy subjects (image size $181 \times 217 \times 181$ and voxel dimension $1.0 \times 1.0 \times 1.0\, \text{mm}^3$). Each LPBA40 subject has expert annotation of 56 localized structures in all four lobes, insula, cingulate gyrus, cerebellum, and brainstem. We refer to (Shattuck et al., 2008) for a complete list of all those structures. We note that, those expert annotations are in no means used as any part of the registration process in this section. They are only used to help evaluate registration accuracies (as detailed in Section 6.2.2.5).

Images in both datasets have been skull-stripped at its release. The only preprocessing we have conducted is N3 bias field correction (Sled et al., 1998). After preprocessing, the bias-field-corrected images are used for the study.

6.2.2.3. Registration Methods As Testbeds

Eight representative registration methods have been included as testbeds: fnirt (Andersson et al., 2008), AIR (Woods et al., 1998), ART (Ardekani et al., 2005), ANTs (Avants et al., 2008), MI-FFD (Rueckert et al., 1999), DROP (Glocker et al., 2008), Demons (Vercauteren et al., 2009), and DRAMMS (Ou et al., 2011). They are included because of being easy to use, being popular in the community, and having demonstrated accuracy and consistency (Klein et al., 2009). Another reason is because they are voxel-wise methods that can readily
take either intensity images or tissue segmentation images without any additional effort. More importantly, they have covered a wide range of techniques for those key components of registration. For instance, they have used different similarity metrics, different transformation models and different optimization strategies. Therefore, testing on them is more likely to show whether tissue segmentation improves registration in general, not just for a specific scenario.

Note that, the focus of this section is not on comparing among those 8 registration methods. Rather, the focus is to test if each one of them will outperform itself when tissue segmentation images are used instead of intensity images. Therefore, parameter setting is less an important issue. As long as we have kept a fixed set of parameters for a specific method, regardless of when it is fed with intensity images or tissue segmentation images, we will fairly serve the purpose of this section. In practice, we have adapted parameter settings from a recent review paper (Klein et al., 2009) for the first 5 methods. The remaining 3 methods are not reported in (Klein et al., 2009), so we set their parameters based on trial-and-error experiments.

We also note that registration is often biased by the selection of template. To avoid any bias in the template selection, we have conducted all possible pair-wise registration within each dataset, by permuting source and template images. This amounts to 240 pair-wise registrations in the NIREP dataset and 1560 pair-wise registrations in the LPBA40 dataset by each of the aforementioned 8 registration methods. On top of that, each pair-wise registration is done in two settings: one using gray-scale intensity images, and the other using automatic tissue segmentation images.

6.2.2.4. Automatic Tissue Segmentation

Automatic tissue segmentation is done using a publically available software tool named FAST (Zhang et al., 2001) (University of Oxford, version 4.1). Use of FAST is also because it is widely-acknowledged as one of the most accurate and consistent brain tissue segmentation
tools in a number of evaluation studies (Bouix et al., 2007; Klausch et al., 2009) in different datasets and by different measurements. In the FAST output, we have kept deterministic tissue segmentation. Further, the segmented three tissue types – CSF, GM, and WM – are re-labeled by intensities 50, 150, 250 to increase contrast among them. Examples of those deterministic tissue segmentation images can be found in the lower left part in Fig. 73.

6.2.2.5. Evaluation of Registration Accuracy

We measure registration accuracy by Jaccard overlap in those 32 (for the NIREP dataset) or 54 (for the LBPA40 dataset) expert-annotated regions. Regional overlap is a widely-used surrogate to indicate registration accuracy (Klein et al., 2009). We note that it can be unreliable, unless regions are highly localized (Rohlfing, 2012). This is understandable, because more localized regions are more likely to approximate landmark points, and therefore their overlaps are more likely to approximate registration accuracy directly measured on landmarks. Fortunately, the expert annotations in both the NIREP and LBPA40 datasets contain highly localized sub-cortical structures. Therefore, Jaccard regional overlap in our study is a fair evaluation of registration accuracy. Specifically, registration of each pair of images will lead to 32 (or 54) Jaccard overlaps in 32 (or 54) regions, and we directly average them into "average Jaccard overlap" to measure accuracy of this registration.

6.2.3. Results

In this section, we will first show and discuss the effect of automatic tissue segmentation on registration accuracy and stability. Then we will measure if the effect is statistically significant.

Effect of Tissue Segmentation on Registration Accuracy and Stability. Fig. 74 shows the accuracies of each registration method with and without using tissue segmen-
tation images. Focus here is not comparing among those testbed registration methods. Rather, focus should be on whether each method outperforms itself when fed with tissue segmentation images instead of intensity images. Two observations can be made:

a) tissue segmentation improves registration accuracy in almost all 8 methods and in both datasets. The only exception is Demons method when used in the NIREP dataset. In both datasets, fnirt method enjoys the biggest increase in overlap-indicated accuracy.

b) more interestingly, tissue segmentation improves stability in almost all 8 registration methods in both datasets. This is reflected by the reduced variation in the average Jaccard overlap. This phenomenon is more evident in fnirt, AIR and DROP.

Let us take a closer look to understand why tissue segmentation improves registration accuracy and stability. Usually tissue segmentation will have two conflicting effects. On the good side, tissue segmentation image suffers less from intensity inhomogeneities compared to intensity image. It defines a sharper boundary among tissue types to better guide registrations at tissue boundaries. On the other side, tissue segmentation also removes contrast within a specific tissue type. For instance, two gray matter structures will exhibit slightly different intensity contrast in intensity images, but the contrast will be completely gone in tissue segmentation image. From the results in Fig. 74, it seems the first effect dominants. Plus, in the second effect, even though intensity contrast is lost, we might still have shape information implicitly from tissue boundary information to guide registration. Therefore, in general tissue segmentation helps registration.

**Statistical Significance of the Effect.** We further tested if the improvement in average Jaccard overlap is statistically significant for those 8 methods. Pair t-test is used, and the resultant $p$ values are reported in Table 9. Quantitative results echo visual findings. Tissue segmentation often leads to significant improvement for registration methods, but not always. It brings more significant improvement to those methods that have smaller Jaccard overlaps when using intensity images alone. fnirt and AIR belong to this category. On the
Figure 74: Box-and-Whisker plots of average Jaccard overlap in the NIREP and LPBA40 datasets. The definition of "average Jaccard overlap" can be found in Section 6.2.2.5.
other hand, some methods are less influenced by introduction of tissue segmentation images. ART, DRAMMS, ANTs and Demons belong to this category. A plausible explanation is that those methods better account for intensity inhomogeneities in themselves, and hence are less benefited from tissue segmentation.

Table 9: Statistical significance of the improved regional overlap. p values are reported in the following table. Significance for Demons in the NIREP dataset is not calculated since overlap decreases with using tissue segmentation images, as can be seen in Fig. 74.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>fnirt</th>
<th>AIR</th>
<th>ANTs</th>
<th>ART</th>
<th>MI-FFD</th>
<th>DROP</th>
<th>Demons</th>
<th>DRAMMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIREP</td>
<td>3.04e-82</td>
<td>2.12e-2</td>
<td>1.15e-7</td>
<td>0.0013</td>
<td>4.75e-83</td>
<td>2.87e-6</td>
<td>—</td>
<td>0.0017</td>
</tr>
<tr>
<td>LPBA40</td>
<td>3.38e-69</td>
<td>1.81e-4</td>
<td>0.077</td>
<td>0.3351</td>
<td>2.70e-12</td>
<td>0</td>
<td>2.93e-4</td>
<td>0.1963</td>
</tr>
</tbody>
</table>

6.2.4. Conclusion

This section examined the effect of automatic tissue segmentation on registration accuracy and stability. To draw a general observation, a large number (28,800) of registrations have been run by using 8 popular registration methods in 2 public brain MRI datasets. Experiment results demonstrate the assumption that tissue segmentation improves registration accuracy and stability, in almost all methods and in both datasets. However, the improvement is not always statistically significant. Those registration methods that are less robust to image inhomogeneities tend to enjoy more significant improvement from automatic tissue segmentation. In the following, we will discuss several important respects in this section and conclude in the end.

For automatic tissue segmentation, we only examined FAST (Zhang et al., 2001) tool. This is because FAST (Zhang et al., 2001) is consistently among the top accurate and stable tissue segmentation methods in recent review studies (Bouix et al., 2007; Klauschen et al., 2009). We can include other tissue segmentation methods in future experiments. And we expect similar effect as found in this section. Also, only deterministic tissue segmentation is used. We leave it to future work to find out whether probabilistic tissue segmentation will also have similar effect.
For measuring registration accuracy, we have used regional overlap in various sub-cortical structures (32 in the NIREP dataset, 54 in the LBPA40 dataset). Those regions are relatively small and highly localized comparing to the whole brain. Therefore, according to (Rohlfing, 2012), regional overlap will serve as a reliable accuracy surrogate in this study. We can use additional metrics such as minimum surface distance in future studies. We expect that additional metrics will perhaps reflect similar results to the regional overlap in our study, as the case in (Klein et al., 2009).

For overall experiment design, we only contrasted between registration methods fed with intensity images and the same registration methods fed with tissue segmentation images. This directly reflects the effect of tissue segmentation. However, we have not studied registration accuracy using intensity image and tissue segmentation image combined. Combining them needs careful design of registration methods, and is therefore outside the scope of this section.

We have attempted to understand the reason behind the improved registration accuracy by tissue segmentation. One plausible explanation is the tissue segmentation suffers less from image inhomogeneities and partial effects than intensity images. It also provides sharper boundaries among tissue types to better guide registration. Although the slight intensity contrast of structures within a specific tissue type may have been lost as a result of tissue segmentation, the sharper boundaries among tissue types seem to implicitly provide shape constraints for better aligning those structures.

In conclusion, experiments in this section showed that, FAST-based automatic tissue segmentation improves registration accuracy and stability in methods that suffer more from image inhomogeneities. However, the effect to DRAMMS is only marginal, partly because DRAMMS has already accounted for image inhomogeneities and contrast differences and has already achieved high registration accuracy.
CHAPTER 7

Discussion and Conclusions

7.1. Summary

This dissertation develops a new medical image registration method named DRAMMS, evaluates its generality, accuracy and robustness in cross-subject and longitudinal registration tasks involving brain, breast and cardiac images, and shows its applications into five studies, which are examples of population, longitudinal and atlas-based segmentation studies. Approaches towards further improving the speed and accuracy of the DRAMMS algorithm are also explored. The proposed algorithm is fully-implemented, extensively-tested and publicly-released, with manuals and tutorials available in the Appendix of this dissertation, to meet the growing needs of software dissemination in clinical and research studies.

The dissertation is motivated by many clinical and research studies that need image registration as a building block; five such examples are mentioned in Chapter 1 which motivate our work. Despite over two decades of extensive research in the field of medical image registration, these studies still pose challenges to many of the existing image registration methods. The first two studies suggest the need for developing population-based statistical atlases of brain lesions (cerebrovascular and cancer, respectively). Understanding the spatial distribution of the lesions, as well as their relationship to clinical variables, is central to these studies. This requires registration of lesion-bearing images from a population of
patients into a common template space, which is often from a normal subject to represent the healthy anatomy. However, lesion-induced missing correspondences present big challenges to image registration. The third study suggests the need for evaluating longitudinal changes in breast cancer patients. Registration of serial scans is central to quantifying disease progression and monitoring treatment effects. However, the large inter-scan change of breast soft tissue and the development of tumor in size, shape and texture, render image registration difficult. The fourth and fifth studies focus on using image registration to propagate anatomical delineations from one or multiple expert-annotated brain or prostate atlases into the target image space. Large anatomical variations among individuals, as well as differences in intensity distributions, image contrast, and fields of view (FOV), especially in multi-site datasets (which are the norms for many clinical studies), make the search for anatomical correspondences a tough task. These studies motivate the need for a new medical image registration method that is

1. generally applicable, i) to different registration tasks (cross-subject, longitudinal), and ii) to images from various organs (brain, breast, heart, prostate, etc);

2. accurate, in establishing anatomical correspondences;

3. robust, with regard to inter-subject anatomical variations, image inhomogeneities, and even missing correspondences caused by pathologies (e.g., lesions and tumors) or field-of-view (FOV) differences.

To meet this goal, the DRAMMS algorithm is developed in Chapter 3, and is extensively validated in Chapter 4. The validated DRAMMS is applied to the five example studies to show its wide application. Chapter 6 explores two attempts towards further improving the speed and accuracy of the DRAMMS framework.

In the following, each chapter is summarized with a listing of results and contributions, which is followed by a statement for future work.
7.2. The DRAMMS Algorithm

Chapter 3 develops the DRAMMS registration algorithm. DRAMMS targets two fundamental problems in most existing registration methods. One problem stems from the difficulty arising in registering images with pathologies (e.g., vascular lesions and tumors). Pathological regions are outliers that do not have clear correspondences in the images of healthy subjects. To reduce their negative impacts to the registration process, we argue that outlier regions, such as pathologies, should be used with low or even zero confidence during registration. The challenge for assigning voxel-wise weights is a need for automatically identifying each voxel’s ability to establish correspondences between images. Ideally, this should be done without segmentations of anatomical structures or regions, to keep the proposed method generally applicable to different registration tasks. The mutual-saliency metric is proposed in this dissertation to overcome this challenge. It quantifies the level of confidence we have in a voxel’s ability to find reliable correspondences. By using the mutual-saliency values to modulate registration, the regions that can establish reliable correspondences are used to drive the registration. Meanwhile, the negative impact of the regions (such as pathologies) whose correspondences are not present in the other image is largely reduced. The second problem is that most registration methods find anatomical correspondences by intensity-based similarity measures. Since intensity alone does not contain anatomical or geometric information, intensity-based registration methods usually encounter ambiguities in establishing voxel-wise correspondences. In the proposed work, more distinctive characterization of voxels is used. The extraction of Gabor attributes and the automatic selection of the optimal components lead to more distinct voxel characterizations and result in more accurate voxel-wise matching. Below is a summary of the contributions in the DRAMMS framework.
7.2.1. Summary of Contributions

1. Attribute matching (including attribute extraction and attribute selection) is developed for finding correspondences more accurately. The selected optimal Gabor attributes render voxels more distinctive and more identifiable in finding correspondences.

2. A mutual-saliency weighting framework is developed for effectively handling the problem of missing correspondences. The mutual-saliency value is automatically and directly derived from the matching of two images. It quantifies the confidence we have in each voxel’s ability to establish reliable correspondences between images. It helps to reduce the negative impact of missing correspondences when they are present.

7.2.2. Future Work

In characterizing the geometric context around each voxel, Gabor attributes are used in the proposed work. The reasons for this choice include its generality (as demonstrated by its successful application in various image analysis tasks (Jain and Farrokhnia, 1991; Zhan and Shen, 2006; Xue et al., 2009; Zhang and Liu, 2004)), and its multi-scale and multi-orientation formulation (which helps render voxels more distinctive). In future studies, other texture attributes can be used, such as wavelet-based attributes (Xue et al., 2004), local frequency attributes (e.g., Liu et al. (2002); Jian et al. (2005)), local intensity histogram attributes (e.g., Shen, 1997; Yang et al., 2008), and many others. Comparison of different texture attributes in the context of image registration is an interesting, yet less explored topic. Experimental comparisons are needed. A theoretical framework to study which texture attributes lead to fewer and smaller scales of local minima may be interesting, too.

The mutual-saliency metric has shown promise to deal with missing correspondences. The
mutual-saliency metric is essentially a measure of the confidence, or the reliability, in the matching between two voxels. There might be alternative ways to define the mutual-saliency metric in the DRAMMS framework. One example is to non-linearly map the calculated mutual-saliency values. Another example is to calculate mutual-saliency values by permutation tests (i.e., to calculate the chance of finding another neighboring candidate that can match with a given voxel with the same level of similarity).

Right now DRAMMS assumes that each subject only has one channel of images in the cross-subject registration task. For example, we register the T1-weighted image of one subject and the T1-weighted image of another subject. This is often called "single-channel" registration. When multi-parametric images are available, we can extend DRAMMS into multi-channel DRAMMS. That is, to register two subjects by simultaneously considering multi-parametric images of each subject (e.g., T1-weighted MRI, T2-weighted MRI, perfusion images, diffusion images, contrast enhancement images, etc). A straightforward extension is to directly stack the attributes from multi-channel data into a unified attribute vector for each voxel. Another extension is in the definition of the mutual-saliency metric. In addition to measuring different confidence levels for different voxels, we can measure different confidence levels for different channels of images. That is, similar to the fact that different voxels have different contributions to the registration process, which is the motivation for the (voxel-wise) mutual-saliency metric, different image channels may also have different contributions to the image registration process. Therefore, one future direction in the multi-channel registration work could be to extend the spatially-varying mutual-saliency metric into a channel-wise and spatial-varying mutual saliency metric. The multi-channel registration is especially useful in monitoring longitudinal change of brain tumors in functional, structural, metabolic, diffusion aspects, as there is recent evidence to show the gain in brain tumor detection rate using multi-parametric (multi-channel) images (Verma et al., 2008).
7.3. Validations of DRAMMS

Chapter 4 presents extensive validations of the proposed DRAMMS algorithm, for demonstrating its generality, accuracy and robustness. Validations are in various registration settings (cross-subject and longitudinal), involving images of various organs (brain, breast and heart). Below is a summary of results and contributions in this chapter.

7.3.1. Summary of Results and Contributions

1. DRAMMS is validated in the cross-subject registration of skull-stripped brain images. DRAMMS is compared with 11 other methods in two public datasets to establish DRAMMS as one of the most accurate methods in the registration of normal, skull-stripped brain images. Contributions include: 1) the inclusion of recently-developed methods (DRAMMS, DROP, Demons) into the evaluation; 2) the study of the correlation between the accuracy and the aggressiveness of deformations; and 3) the presentation of two sets of parameters to explore a registration method’s stability with regard to parameter changes.

2. DRAMMS is validated in the cross-subject registration of raw brain images. Compared with Demons, ANTs and fnirt registration methods, DRAMMS is found to have the highest accuracy, in two out of three datasets, in aligning brain masks (ANTS being most accurate in the other dataset). DRAMMS is found to be more robust than three other methods (ANTS, Demons, fnirt), with regard to image inhomogeneities, background noises, and even field-of-view differences. The advantage of DRAMMS over the other three methods is larger in multi-site datasets, which are the norms for many clinical studies.

3. DRAMMS is validated in the registration of lesion-bearing brain images with a nor-
mal brain template. Simulated lesions of increasing sizes are used to measure the robustness of a registration method with regard to the presence of vascular lesions. DROP and DRAMMS are found to have the highest levels of robustness compared with fnirt, ART, ANTs, and the cost-function-masking variants of fnirt and ANTs.

4. DRAMMS is validated in the registration of brain images with tumor recurrences to a normal brain template. The accuracy is measured against expert-annotated tumor recurrence regions and anatomic landmarks. Among the four methods compared (ANTs, fnirt, Demons, DRAMMS), ANTs is found to align tumor recurrence regions with the highest accuracy, whereas DRAMMS is found to have the highest accuracy in registering regions close to and far from tumor recurrence regions.

5. DRAMMS is validated in the cross-subject registration of cardiac images (all non-cardiac structures being removed). DRAMMS is found to have the highest accuracy among 12 registration methods in aligning left ventricles, right ventricles and myocardium regions of the heart.

6. DRAMMS is validated in the registration of longitudinal breast cancer images. Based on landmarks independently defined by two raters, DRAMMS produce the smallest landmark errors compared with two other methods (flirt and ART). The errors between DRAMMS and raters are found to be comparable to the inter-rater landmark errors. This is perhaps the first study to quantify inter-rater and the algorithm-to-rater errors in different regions (tumor regions versus normal regions), and in different patient groups (those who show complete response to chemotherapy and those who do not) of breast cancer patients.
7.3.2. Future Work

Future validations include evaluating DRAMMS and the other methods in multi-modality registration tasks, such as the task of registering MRI and CT images of the same subject, or more interestingly, the task of registering functional data (e.g., fMRI, PET) and structural data (e.g., MRI) of the same subject. In addition, the stability of a registration method with regard to parameter changes is another area that needs further investigation.

7.4. Applications of DRAMMS

To show the wide application of DRAMMS, Chapter 5 applies the DRAMMS method to five clinical and research studies. They are examples of population studies, longitudinal studies and atlas-based segmentation studies. Below is a summary of results and contributions in this chapter.

7.4.1. Summary of Results and Contributions

1. DRAMMS is applied to constructing statistical atlases of brain vascular lesions and tumors. The constructed atlases show the frequent locations for white matter lesion occurrence in the Type II diabetic population. They also lead to correlations of the spatial distributions of pathologies with clinical variables such as gender and the disease duration.

2. DRAMMS is applied to quantifying, at the voxel level, the longitudinal change of breast cancers. This is an important step towards evaluating the effect of chemotherapy.
3. DRAMMS is applied to segmenting the brain and the prostate from the raw brain and raw prostate images in multi-atlas-based frameworks. In the brain extraction, DRAMMS-based multi-atlas segmentation pipeline scores higher accuracies than a state-of-the-art brain extraction method ROBEX (Iglesias et al., 2011), with a 0.97–0.98 average dice overlap with expert annotations. In the segmentation of the prostate, the multi-atlas pipeline scores a 0.84 average dice overlap with expert annotations in a multi-site dataset, which is comparable to the accuracy in a single-site dataset as reported in (Klein et al., 2008).

7.4.2. Future Work

Results in this chapter are mostly preliminary. Therefore, future work will be to continue the five studies in this chapter. For example, in the brain tumor studies, we need to collect images (both pre-resection and post-resection with tumor recurrences), from more patients. This way, we can construct tumor recurrence atlases that more accurately approach the actual distributions of tumors in the brain tumor population. The constructed atlases of tumor recurrences will be correlated with the atlases of the initial tumor occurrences in the same population. This will enable a study of whether tumor recurrence follows certain patterns (e.g., the distance and the scale with regard to the initial occurrences; whether recurrences follow main vessel networks or white matter connectivity pathways, etc). In the breast cancer study, our preliminary results in this chapter have shown that DRAMMS can quantify tumor changes in breast cancer patients. Future work will investigate the relationship between the patterns of the tumor changes (e.g., the change rate, intensity residuals) with the image appearances in the patients’ baseline images. This will help us adapt treatment methods to each individual patient by using the baseline image appearances as predictors of response.

Future work also includes applying DRAMMS to other population studies and longitudinal
studies. An example population study is the investigation of Alzheimer’s Diseases (AD) patients. DRAMMS can be used to normalize brain images from AD patients into the same template space, based on which the AD-related imaging bio-markers can be studied. An example longitudinal study is to investigate the growth of the mouse brain, or the growth of the human pediatric brain.

7.5. Towards Improving DRAMMS Speed and Accuracy

Chapter 6 explores two attempts to further improve the speed and accuracy of the DRAMMS framework. Below is a summary of results and contributions in this chapter.

7.5.1. Summary of Results and Contributions

1. A discreteDRAMMS framework is developed to reduce the computational cost of DRAMMS. The discreteDRAMMS framework preserves the attribute matching and mutual-saliency components in the DRAMMS framework. In contrast to DRAMMS, which finds correspondences at each and every voxel, the proposed discreteDRAMMS establishes correspondences at a number of automatically-located landmark voxels. A forward-backward one-to-one matching criterion is designed to ensure that the extracted landmark correspondences inherit the accuracy and robustness from the original DRAMMS framework. A spatial clustering method is developed to encourage landmarks to be spatially dispersed in the image space, so that they can be used to guide voxel-wise registration in various image regions.

2. A study is carried out to test whether the automatic segmentation of brain tissue types can further improve the accuracy in cross-subject brain registration tasks. Results show that the improvement of the registration accuracy is marginal in methods such
as DRAMMS, ANTs and Demons, which have already obtained high accuracies in cross-subject registration of brain images.

7.5.2. Future Work

Future work will focus on improving the speed and accuracy of the DRAMMS framework. The speedup of DRAMMS is important to satisfy the needs in many large-scale clinical studies that require registration of hundreds or even thousands of images. Two paths can be taken to improve the speed. One path is related to the algorithm design. The discrete-DRAMMS developed in this chapter follows this path. The automatically-located landmark pairs in the discreteDRAMMS framework can be used to guide other intensity-based voxel-wise methods. Therefore, we can achieve relatively high accuracies, by utilizing DRAMMS’ robustness, and relatively fast computations, as intensity-based voxel-wise methods are computationally more efficient than DRAMMS. The second path towards improving the speed of DRAMMS is to use more advanced hardware, such as the Graphics Processing Unit (GPU). The DRAMMS framework calculates attributes and mutual-saliency values independently at each voxel. Therefore, the calculations at all image voxels can be parallelized and re-programmed into multiple threads, which is suitable for GPU processors.

Right now, DRAMMS is a general-purpose image registration method that does not require any task-specific prior knowledge. To improve the accuracy of DRAMMS, we can incorporate prior knowledge into DRAMMS for a specific task. The prior knowledge can be in the form of structure masks, which constrain the DRAMMS framework in specific regions outlined by the structure masks. For example, registration of raw cardiac images suffers from a high failure rate because non-cardiac structures, which dominate the image space, differ greatly among subjects. The mask of the heart, which provides prior knowledge (location, size, orientation, etc) of the heart, can be used to constrain DRAMMS to establish correspondences only within the heart region, thereby increasing the success rate.
7.6. Concluding Remarks

Deformable registration is a fundamental problem in medical image computing, and has been at the forefront of research in the field of medical image analysis for over two decades. It is central in analytic methods for understanding and quantifying population trends of imaging phenotypes, for following longitudinal disease progression or evaluating treatment effects, for fusing the complementary information from multi-modality imaging data, and for capturing structural-functional correlations.

This dissertation makes contributions to the field of medical image registration, by developing a novel deformable image registration algorithm termed DRAMMS, which is short for "Deformable Registration via Attribute Matching and Mutual-Saliency Weighting". DRAMMS targets two fundamental challenges in most existing image registration methods — how to accurately establish correspondences, and how to effectively handle missing correspondences, which arise in the presence of pathologies or highly variable anatomical content. In particular, DRAMMS establishes correspondences based on high-dimensional attributes at each voxel. An attribute extraction and attribute selection pipeline is developed. It leads to characterizations of the geometric/anatomical context in the vicinity of each voxel, leading to more accurate voxel matching compared with the voxel matching based on the intensity information alone. DRAMMS handles the missing correspondences by a newly-developed mutual-saliency metric. The mutual-saliency value indicates whether a voxel can establish reliable correspondences between images. It is automatically computed from the matching of two images, and is iteratively updated during the registration process. It helps drive the registration with regions that can establish correspondences in the other image, and reduce the negative impact from regions (such as pathologies) that
cannot establish correspondences in the other image.

The proposed DRAMMS framework is extensively validated in this dissertation for various registration tasks (cross-subject, longitudinal) involving images of various organs (brain, heart and breast). Validations establish the generality, accuracy and robustness of the DRAMMS algorithm. The validated DRAMMS algorithm is then applied to five studies, which are examples of population studies, longitudinal studies and atlas-based segmentations of anatomical structures. Finally, to meet the growing needs in multi-site, large-scale research and clinical studies, DRAMMS is fully-implemented, extensively-tested and publicly-released to the medical image analysis community.

Future work will focus on evaluating DRAMMS in more registration tasks (e.g., multimodality data), further improving the speed and accuracy of DRAMMS, and applying DRAMMS to large-scale and multi-site clinical and research studies.
APPENDIX A

DRAMMS Software Package

This appendix introduces the DRAMMS software. After more than 3 years of work by three developers (Yangming Ou, Andreas Schuh and Aristeidis Sotiras), with more than 1800 revisions and 8 major internal tests (as of December 3, 2012), and with the help of almost 20 constructive testers here in the lab and worldwide, DRAMMS is a now an extensively-tested and publicly-available software package free to the academic community (http://www.rad.upenn.edu/sbia/software/dramms/). Fig. 75 shows the screenshot of the first page in DRAMMS website.

In the following, Section A.1 introduces the DRAMMS software, followed by download and installation instructions in Sections A.2 and A.3. Then, two major sections — A.4 and A.5 — describe the DRAMMS manual (including the basic tool for image registration and the auxiliary tools for deformation manipulation) and DRAMMS tutorials (the example usages, expected results and other options in 12 typical registration tasks involving cross-subject, longitudinal and multi-modality registrations in 2D-to-2D or 3D-to-3D brain, cardiac, breast and prostate images). Section A.6 answers a few frequently asked questions. Finally, Sections A.7 and A.8 present DRAMMS-related publications and acknowledge people who have contributed to the development of the DRAMMS software.
A.1. Introduction to the DRAMMS Software

Deformable Registration via Attribute Matching and Mutual-Saliency Weighting (DRAMMS) ([Ou et al. 2011](http://www.rad.upenn.edu/sbia/software/dramms/)), is a software package designed for deformable 2D-to-2D and 3D-to-3D image registration.

DRAMMS can be used in many registration tasks, including:

- **Cross-subject** registration (can be brain images, breast images, cardiac images, etc);
- **Mono-** and **Multi-modality** registration (MRI, CT, histology);
- **Longitudinal** registration (pediatric brain growth, cancer development, etc);
- Registration under **missing correspondences** (small lesions, tumors, histological cuts).

DRAMMS is implemented as a Unix command-line tool. It is fully automatic and easy to use — users input two images, and DRAMMS will output the registered image and deformation. No need for pre-segmentation of any structures, no need for any prior knowledge, and no need for human initialization or intervention.
DRAMMS is implemented as a Unix command-line tool. It is fully automatic and straightforward to use — users input two images, and DRAMMS will output the registered image and deformation. No need for pre-segmentation of any structures, no need for any prior knowledge, and no need for human initialization or intervention.

### A.2. Software Download

This sub-section describes the download of the DRAMMS software. Before downloading, we first describe the license issues and the system requirement.

#### A.2.1. Software License

Most parts of DRAMMS software (attribute extraction, attribute matching, mutual-saliency matching, deformation operations) are freely available under a BSD-style open source license that is compatible with the Open Source Definition by [The Open Source Initiative](https://opensource.org/licenses/BSD-2-Clause) and contains no restrictions on use of the software. The full SBIA license text is included with the distribution package and available online.

The Fast-PD optimization part within DRAMMS software is a modified version of the original [FastPD](https://bitbucket.org/members/fastpd). The modification of FastPD and the release of modified FastPD source code within DRAMMS software are approved by the FastPD’s owner under a written authorization for research and academic purpose only. Note that the FastPD optimization package is protected from several international pending patent applications. For commercial use of this FastPD optimization package (both modified and original versions), please contact [here](mailto:support@drumms.org).
The affine part within DRAMMS is handled by flirt tool from FSL software package, therefore bears FSL license that grants non-commercial use only.

A.2.2. Documentation

DRAMMS Flyer (2 pages, 0.4MB): A quick overview of DRAMMS and its use.
http://www.rad.upenn.edu/sbia/software/dramms/_downloads/DRAMMS_Flyer.pdf

http://www.rad.upenn.edu/sbia/software/dramms/manual.html

DRAMMS ChangeLog Summary of changes, new features, and bug fixes.
http://www.rad.upenn.edu/sbia/software/dramms/changelog.html

A.2.3. System Requirements

Operating System: Linux, Mac OS X

Memory Requirement: DRAMMS requires a considerable amount of memory. The exact memory requirement depends on the dimensions of the images. But generally, the default use of DRAMMS should ‘not’ consume more than 12GB memory even when the images are large (e.g., 1024*1024*600). The memory consumption for some typical image sizes are:

- ~0.5GB for a typical pair of 2D images (e.g., 256*256),
- ~2.5GB for a typical pair of 3D cardiac/breast MR images (e.g., 256*256*100),
- ~3.0GB for a typical pair of 3D brain MR images (small) (e.g., 256*256*124),
- 4-10GB for a typical pair of 3D brain MR images (big) (e.g., 256*256*256).
• 10-11GB for a typical pair of 3D head+neck CT images (big) (e.g., 512*512*350).

The good thing is that users can choose to use less memory if their system can only afford less (see FAQ.4 in Section A.6.4). Be aware, though, that this may cause a slight decrease in registration accuracy.

A.2.4. Register for Download

Please register here (http://www.rad.upenn.edu/sbia/software/request.php?software=dramms) to receive an email with the download links of the software.

A.3. Software Installation

This sub-section contains step-to-step instructions to install DRAMMS software.

A.3.1. Prerequisites

<table>
<thead>
<tr>
<th>Dependency</th>
<th>Version</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMake</td>
<td>2.8.4</td>
<td>Needed to compile and build DRAMMS.</td>
</tr>
<tr>
<td>FSL</td>
<td>4.1.5</td>
<td>For affine registration in DRAMMS.</td>
</tr>
<tr>
<td>NiftiClib</td>
<td>2.0.0</td>
<td>Nifti library to support nifti I/O.</td>
</tr>
<tr>
<td>BASIS</td>
<td>2.1.0</td>
<td>A SBIA meta-project to standardize software development.</td>
</tr>
<tr>
<td>FastPD</td>
<td></td>
<td>Optimizer for DRAMMS.</td>
</tr>
</tbody>
</table>

Out of these five dependencies:

users need to install two dependencies — CMake and FSL — before DRAMMS
users need not to install the other three dependencies — BASIS, NiftiCLib and FastPD are included in the DRAMMS source package and by default built as part of the installation.

A.3.2. Build and Installation

Please follow commands below in a shell/terminal (e.g., [Bash](https://www.gnu.org/software/bash/)). They will configure and build DRAMMS using [GNU Make](https://www.gnu.org/software/make/). The main CMake configuration file (CMakeLists.txt) is located in the `dramms-1.1.0-source/build/` subdirectory.

Step 1. Extract source files:

`tar xzf dramms-1.1.0-source.tar.gz`

Step 2. Change to build directory:

`cd dramms-1.1.0-source/build`

Step 3. Run CMake to configure the build tree:

`ccmake .`

After execution of this command, you will see a screen like below (Fig. ccmake).

In this ccmake interface, please do:

3.1. Change `CMAKE_INSTALL_PREFIX` to the folder you want to install DRAMMS into. Make sure you have the write access to this directory.

3.2. Keep pressing `c` until option `g` is available.

3.3. Then press `g` to generate makefiles and quit this ccmake window.
Step 4. Build and install DRAMMS:

```
make
```

After the above compilation and build finished successfully, DRAMMS is installed into the directory specified by the `CMAKE_INSTALL_PREFIX` (set during build configuration in step 3). The DRAMMS Manual are located in the `bin/` subdirectory.

### A.4. Software Manual

This sub-section introduces tools in DRAMMS software package. Two major parts are presented: a main command-line tool for image registration, and a pool of tools for manipulating the obtained deformation (to warp image, to calculate Jacobian map, to calculate RAVENS map, to add/compose/average/subtract two deformations, etc).
A.4.1. Main Command for Image Registration

The main command of DRAMMS which registers two images and optionally performs further analysis of the obtained deformation is named *dramms*.

```
dramms --source sourceimage.hdr --target targetimage.nii
     --outimg outimage.img     --outdef outdef.nii.gz
```

**Supported File Formats:** Recommend NIfTI-1 (.hdr+.img; .nii; .nii.gz); also accept ANALYZE-7.5 (.hdr+.img).

**Supported Datatypes:** byte (unsigned char, uint8), int8, short, int16, uint16, float, float32, int32.

**Parameter Settings:** The default settings of *dramms* will give reasonable results in many cases. We recommend a look at the [Tutorials](#) Section A.5 for task-specific uses. If in a particular case neither the default nor the example settings give satisfactory results, please refer to [FAQ.1](#) in Section A.6.1 for the most important parameters to tune.

**Memory Settings:** Please keep in mind that *dramms* may terminate due to short of memory. Memory requirement is largely dependent on image size and specific task — see [System Requirement](#) in Section A.2.3 for general rules, and the [Tutorials](#) in Section A.5 for task-specific examples. How to reserve memory can be found in [FAQ.3](#) in Section A.6.3. If you can only afford less memory, please refer to [FAQ.4](#) in Section A.6.4 for how to reduce memory usage in *dramms*.
## A.4.2. Auxiliary Commands for Deformation Operations

<table>
<thead>
<tr>
<th>Operation</th>
<th>Command</th>
</tr>
</thead>
<tbody>
<tr>
<td>Warp images</td>
<td><code>dramms-warp</code></td>
</tr>
<tr>
<td>Calculate Jacobian Determinants</td>
<td><code>dramms-jacobian</code></td>
</tr>
<tr>
<td>Calculate RAVENS maps</td>
<td><code>dramms-ravens</code></td>
</tr>
<tr>
<td>Invert/Smooth deformation, or Print-out displacement at a voxel</td>
<td><code>dramms-defop</code></td>
</tr>
<tr>
<td>Compose/Add/Subtract/Average two transformations</td>
<td><code>dramms-combine</code></td>
</tr>
</tbody>
</table>

### A.4.2.1. Warp Images

There are two ways to warp an image – a stand-alone command, or as part of the main `dramms` script.

**Option 1: stand-alone command**  The first option is via a stand-alone command named `dramms-warp`, which warps an input image from source image space to target image space, using the deformation obtained when registering source to target. We recommend this option if you have already obtained a deformation.

For trilinear interpolation:

```
dramms-warp inputImage.hdr def.nii.gz warpedImage.img
```

For nearest neighbor interpolation:

```
dramms-warp inputImage.hdr def.nii.gz warpedImage.img -n
```

**Option 2: using dramms**  The second option is suitable if you have no deformation yet, but want to warp another image (i.e., not the source image used to obtain the deformation) from source to target space during (or more precisely, right after) calculating deformation.
This is for example useful in atlas-based segmentation or in structure-based functional mapping.

In this case, use the -W option of dramms:

```
dramms -S A.hdr -T B.hdr
-0 A2B.hdr -D def.nii.gz
-L Alabel.hdr -W Alabel_warped.hdr -r 0
```

In above example, dramms calculates the deformation from image A to image B and then uses it to warp the image Alabel (usually ROI labels or a functional image) from A into B space, using nearest-neighbor (-r 0). For trilinear interpolation, use -r 1 instead.

### A.4.2.2. Calculate Jacobian Determinant Map

Jacobian determinate is an indicator for volumetric change at each voxel. It encodes how each voxel has deformed in term of its volumetric change. Its value is greater than 1 if there is volume expansion, 1 if volume preservation, and less than 1 if volume contraction, 0 if volume vanish, and less than 0 if this region/voxel has been deformed into other regions/voxels, which is usually undesirable.

There are two ways to calculate Jacobian Determinant of a deformation – a stand-alone command, or as part of the main dramms script.

**Option 1: stand-alone command**  The first option is via a stand-alone command named `dramms-jacobian`. We recommend this option if you have already obtained a deformation.

```
dramms-jacobian def.nii.gz outputJacDet.hdr
```

**Option 2: using dramms** The second option is suitable if you have no deformation yet, but want to output the Jacobian Determinant map during (or more precisely, right after) calculating the deformation.
In this case, use the `-J` option of `dramms`:

```
dramms -S A.hdr -T B.hdr
    -O A2B.hdr -D def.nii.gz
    -J jacdet.hdr
```

### A.4.2.3. Calculate RAVENS Tissue Density Map

RAVENS maps are tissue density maps. They are always defined in target space. A RAVENS value at each voxel encodes how many times of volume has been deformed from source image to target image at the current voxel. In other words, RAVENS value records *volumetric change ratio* (subject:template).

The differences from Jacobian Determinants are:

1. RAVENS values are never negative, regardless of the existence of deformation self-foldings; therefore
2. unlike Jacobian Determinant maps, RAVENS maps are volume-preserving; and also
3. RAVENS maps are localized – we usually calculate one RAVENS map per ROI (hence, RAVENS calculation needs ROI masks, even though calculating deformation does not need ROI masks).

There are two ways to calculate RAVENS maps – a stand-alone command, or as part of the main `dramms` script.

**Option 1: stand-alone command** The first option is via a stand-alone command named `dramms-ravens`. We recommend this option if you have already obtained a deformation.

```
dramms-ravens $\{\text{labelImage}\} \{\text{templateImage}\} \{\text{DefField}\}
    \{\text{RAVENS\_prefix}\} -m <\text{int}>[,<\text{int}>,...]
```
Required arguments:

- **labellImage**: label image in subject space.
- **templateImage**: template image, where RAVENS map will reside.
- **DefField**: deformation field (generated when registering template to subject image).
- **RAVENS_prefix**: prefix for all RAVENS maps.
- **-m <int>,<int>,...**: labels of up to 5 ROIs where RAVENS maps will be calculated (we will output one RAVENS map for each ROI).

Optional arguments:

- **-f <int>**: scale factor (default: 1000)
- **-h**: help; usage of this program.

**Option 2: using dramms** The second option is suitable if you have no deformation yet, but want to output RAVENS maps during (or more precisely, right after) calculating deformation.

In this case, RAVENS maps can be calculated using the **-R**, **-L** and **-l** options of **dramms**:

```
dramms -S subject.hdr -T template.hdr
         -O temp2subj.hdr -D def_temp2subj.nii.gz
         -R prefix_ravens_inTemplateSpace
         -L labelsubj.hdr -l 10,150,250
```

- The **-R** argument specifies the prefix of the output RAVENS maps. The eventual outputs will be `$\{prefix\}_$\{label1\}.nii.gz$, `$\{prefix\}_$\{label2\}.nii.gz$, ..., where `$\{label1\}$`, `$\{label2\}$`, ..., are given by the **-l** arguments as explained below.

- The **-L** argument specifies the input label image or ROI image in subject image space.
For example, -L labelsubj.hdr inputs a label image where, say, white matter region is labeled as 250, gray matter is labeled as 150, and CSF as 10).

- The -l arguments specify the labels of the regions in which RAVENS maps will be calculated. For example, -l 10,150,150 results in the computation of three RAVENS maps, one for each region labeled by 10, 150, 250 in labelsubj.hdr. Up to 5 labels can be specified at a time.

Fig. 77 shows input and output results in this example. With this command, there will be 3 output RAVENS maps in template space:

```
prefix_ravens_inTemplateSpace_10.nii.gz,
prefix_ravens_inTemplateSpace_150.nii.gz,
prefix_ravens_inTemplateSpace_250.nii.gz.
```

1. RAVENS calculation requires label image, because RAVENS is usually region/structure-specific.

2. RAVENS values are multiplied by a factor (default 1000) to enhance contrast.

3. RAVENS maps are saved in signed short datatype in template image space. But the registered image in this case is in subject space. See figure above.

**A.4.2.4. Operations on Deformation**

**Read Displacement At A Voxel** To display the displacement of point (x,y,z) in deformation def.nii.gz in the command window, use the following

```
dramms-defop -c ${x},${y},${z} def.nii.gz
```

**Invert A Transformation** Invert an affine matrix:
Figure 77: An example of how to use *dramms* main script to calculate RAVENS maps.
Invert a deformation:

```
dramms-defop -i affine.mat inverted_affine.mat
```

**Smooth A Deformation** To let the output deformation file replace the input deformation file:

```
dramms-defop -s def.nii.gz
```

To keep input deformation and generate a new smoothed deformation:

```
dramms-defop -s def.nii.gz smoothed_def.nii.gz
```

### A.4.2.5. Combine Two Transformations

**Concatenating** two deformations (one A->B and the other B->C) will lead to deformation A->C:

```
dramms-combine -c in_def_A2B.nii.gz in_def_B2C.nii.gz out_def_A2C.nii.gz
```

**Concatenating** affine transformation (A->B) and deformation B->C will similarly produce deformation A->C:

```
dramms-combine -c -f A.nii.gz -f B.nii.gz
in_affine_A2B.mat in_def_B2C.nii.gz out_def_A2C.nii.gz
```

**Add** two deformations which are obtained from the same source and the same target space:

```
dramms-combine -a in_def_1.nii.gz in_def_2.nii.gz out_def_1add2.nii.gz
```

**Subtract** two deformations which are obtained from the same source and the same target space:
A.5. Software Tutorials

This tutorials sub-section is a major section in this software appendix. The tutorials include 12 typical examples covering a wide variety of image registration tasks (2D-to-2D and 3D-to-3D cross-subject, longitudinal and multi-modality registrations of brain, breast, cardiac and prostate images). Fig. [78] is a summary of all 12 examples in this tutorial sub-section. Each example will be presented with DRAMMS command, DRAMMS results, computational resources needed and alternative options if different results are needed.

A.5.1. 2D Images

A.5.1.1. Tutorial 1: 2D Simulated Images

Introduction This is like a “Hello World!” example of DRAMMS, to taste a little bit flavor of how DRAMMS works, and how its result/deformation looks like.

Result

Command

```
dramms -S source.hdr -T target.hdr
```

Average two deformations which are obtained from the same source and the same target space:

```
dramms-combine -s in_def_1.nii.gz in_def_2.nii.gz out_def_1minus2.nii.gz
```

```
dramms-combine -m in_def_1.nii.gz in_def_2.nii.gz out_def_mean.nii.gz
```
Figure 78: A list of 12 examples included in this tutorial sub-section. They are detailed in Sections A.5.1.1 — A.5.5.1.
Figure 79: Registration of two 2D simulated images.

-0 S2T.hdr -D def_S2T.nii.gz

Resources Needed  Registering this pair of 2D images (256 x 256) takes 42MB memory and finishes in 1.5 min in Linux OS (2.80GHz CPU).

A.5.1.2. Tutorial 2: 2D Histological and MR Images of The Same Mouse Brain

Figure 80: Registration of two 2D multi-modality (histology and MR) slices of a mouse brain.

Result

Command

dramms -S src_2DHist.nii.gz -T trg_2DMRI.nii.gz

-0 src2trg.nii.gz -D def_src2trg.nii.gz -x 7

Here, we have used option -x 7 to place control points every 7 pixels in x and y directions. If this option is not used, the program will place control points every \( \text{int}(\frac{\text{sizeX}}{25}) \) (=10 in this 256 x 256 image) pixels, which was the case in another example regarding
2D simulated image. Placing denser control points in this example enables the program to capture more local deformations.

**Resources Needed**  Registering this pair of 2D images (256 x 256) takes 36MB memory and finishes in 1.3 min in Linux OS (2.80GHz CPU). Compared to 1.5 min computational time in example of 2D simulated images, denser control point grids usually takes less computational time.

### A.5.2. 3D Brain Images

#### A.5.2.1. Tutorial 3: Brain MRI of Different Subjects (Skull-stripped)

**Introduction**  Registering skull-stripped and healthy brain images across subjects is a classic problem in image registration. It applies to brain atlas construction, atlas-based structure/ROI labeling. It is also one of the first steps for sub-population differentiation (e.g., normal v.s. dementia populations).

**Result**  A typical set of DRAMMS registration results can be found in Fig. 28 in Chapter 4. Below is the repetition of that same figure.

**Command**

```
dramms -S src_brain1.nii.gz -T trg_brain2.nii.gz
-0 src2trg.nii.gz -D def_src2trg.nii.gz
```

**Other Options**  Usually, the default parameter setting will give reasonable results in cross-subject brain MR registrations.

If one wants to make deformation smoother (often less accurate) or more aggressive, one can look at smoothness weight, -g option. This weight is usually between 0 to 1, higher
Figure 81: Registration of 3D brain MR images from two different subjects (skull-stripped) in NIREP dataset.
weight for smoother deformation. The default weight is 0.2 (i.e., -g 0.2).

If one wants to recover larger or smaller deformations than default parameters, please look at -x, -y and/or -z options. They are directly related to search range in x, y, z directions.

Resources Needed  Registering this pair of 3D images (target image space 256 x 256 x 124 voxels, 1.0 x 1.0 x 1.0 mm^3/voxel) takes 5.9 GB memory and finishes in 38 minutes in Linux OS (2.80GHz CPU).

If one has less memory to offer, please look at option -u to reduce the memory usage to several other levels (the lowest level being 1.6GB in this pair of images).

A rule of thumb is that, computational time and memory usage go up when image size increases: time is roughly linear to image size, and memory is roughly linear to the square root of image size.

A.5.2.2. Tutorial 4: Raw Brain MR Images (with Skull, Different FOV and Background Noise)

Introduction  Registering raw images is a much more difficult problem than registering brain images without skull or background noise. Yet it is useful for propagating of brain masks, for aligning non-brain structures, and is needed in cases skull-stripping may fail.

Major challenges include: 1) large inter-subject variations; 2) large amount of background noise and inhomogeneity during image scanning; 3) different fields-of-view (FOVs), where one image may contain more contents than the other image, like the example shown below.

Result  The following results are repetition of the results already shown in Figs. 37 and 39 in Chapter 4. They are DRAMMS registration results in within-dataset registration and across-dataset registration, respectively.
Figure 82: Typical DRAMMS registration results in within-dataset registrations. Two subjects are both from OASIS dataset and do not present large FOV difference. DRAMMS well handles the differences between intensity, contrast and anatomy.

Figure 83: Typical DRAMMS registration results in across-dataset registrations. The source image is from OASIS dataset and the target image is from ADNI dataset. DRAMMS well handles the differences between intensity, contrast and anatomy, and especially the large difference in FOVs, which is a common yet fairly challenging problem in raw brain image registrations.
Command

```
dramms -S src_rawbrain1.nii.gz -T trg_rawbrain2.nii.gz
    -O src2trg.nii.gz        -D def_src2trg.nii.gz
```

Other Options  If two raw images with skulls are from the same dataset, the default parameter setting will usually give reasonable results.

If two raw images with skulls are from different datasets, and/or from different institutions, the default parameters still work successfully in a majority of the cases (>85% in our 300+ tests), just like the figure shown above.

In extreme cases (when FOV are too different, when background noise are too much, when inter-subject difference is too large), the default parameters may fail. In those extreme cases, we can make DRAMMS work again in most cases by enlarging search range, i.e., setting bigger values in \(-x\), \(-y\), \(-z\) options in dramms command. A rule of thumb is to set \(-x\) value to be \(\text{int(imageSizeX}/35)\), \(-y\) value to be \(\text{int(imageSizeY}/35)\) and \(-z\) value to be \(\text{int(imageSizeZ}/40)\). For example, if the target image is 256 x 180 x 256, then one can first try all settings default:

```
dramms -S source.nii.gz -T target.nii.gz
    -O S2T.nii.gz
```

if in an extreme situation the default setting fails, one can most likely make DRAMMS work again by enlaring search range like below:

```
dramms -S source.nii.gz -T target.nii.gz
    -O S2T.nii.gz
    -x 7    -y 5    -z 6
```

Here the \(-x\), \(-y\) and \(-z\) values are determined according to the above mentioned rule of the thumb.
Resources Needed  Registering this pair of 3D images (target image space 180 x 256 x 256 voxels, 1.20 x 0.94 x 0.94 mm^3/voxel) takes 8.7 GB memory and finishes in 87 minutes in Linux OS (2.80GHz CPU).

If one can afford less memory, please use -u option to choose memory usage in different levels (the lowest being about 1/4 of maximum memory used). This may however slightly reduce registration accuracy.

A.5.2.3. Tutorial 5: Longitudinal Pediatric Images

Introduction  We often use registration to follow brain development in pediatrics. Or, we use registration to propagate the skull-stripping and tissue/structure segmentation from one time point to another time point.

In such registrations, the main difficulties are the large structure change over time, and even development of new structures.

Result

Command

```
dramms -S src Olderage.hdr -T trg Youngerage.hdr
-O src2trg.nii.gz -D def S2T.nii.gz
-g 0.2
```

Other Options  If the structural change is too large, one can increase search range in DRAMMS by setting control point spacing via -x, -y and/or -z options.

Resources Needed  Registering this pair of 3D images (target image 256 x 256 x 175 voxels, 1.0 x 1.0 x 1.0 mm^3/voxel) takes 8.2 GB memory and finishes in 30.5 minutes in Linux OS (2.80GHz CPU).
Figure 84: Registration of longitudinal images of a same baby, to follow brain development.
If one can afford less memory, please use \(-u\) option to choose memory usage in different levels (the lowest being about 1/4 of maximum memory used). This may however slightly reduce registration accuracy.

**A.5.2.4. Tutorial 6: Lesion Brain Onto Normal Brain Template**

**Introduction**  Registration from lesion brain to normal template can help us normalize a population of brains, all having lesions, onto a same template. Then we can observe lesion distributions, correlate lesion occurrence with functional deficit, and monitor treatment effect on slowing or accelerating lesion growth.

The main difficulty is the lesion part - it is present in lesion brain but not normal template. The mutual-saliency component in DRAMMS helps alleviate this problem.

**Result**

**Command**

```
dramms -S src_LesionedBrain.hdr -T trg_NormalTemplate.hdr
    -O src2trg.nii.gz -D def_S2T.nii.gz
    -c 2
```

Here \(-c \ 2\) options turns on (and save) the mutual-saliency weighting part of DRAMMS. Mutual-saliency weight is an automatically derived value that weights each voxel based on its ability to establish reliable correspondences. It is helpful in our case, since mutual-saliency map will use lesions regions with less confidence and rely on other normal regions to drive the registration. Here, no segmentation or prior knowledge of lesion regions is needed. There is no need for segmentation lesion, or any prior knowledge of lesion size or shape.
Figure 85: Registration of a 3D brain MR images with lesions (source) onto a healthy brain template (target). Red arrows point out lesions in source image (a) and in registered image (c). Because of mutual-saliency weighting, the negative impact of lesion is reduced. As a result, the surrounding normal regions are well aligned, and the lesion region is mapped to visually the right place.
**Resources Needed**  Registering this pair of 3D images (target image size 256 x 256 x 198 voxels, 1 x 1 x 1 mm$^3$/voxel) takes 9.5 GB memory and finishes in 62 minutes in Linux OS (2.80GHz CPU).

If one can afford less memory, please use `-u` option to choose memory usage in different levels (the lowest being about 1/4 of maximum memory used). This may however slightly reduce registration accuracy.

**A.5.2.5. Tutorial 7: Brain with Tumor Recurrence Onto Normal Brain Template**

**Introduction**  Here we present an example of using DRAMMS to register an image with recurring brain tumor to a healthy brain template. This will help us correlate tumor recurrence and its original occurrence in normal template space. Another application is to do this for a population of brain tumor subjects, to observe which region is more likely to have recurrence, whether recurrence follows connectivity or follows vessels in population study.

The main difficulties are missing correspondences:

1. after tumor resection, there is usually a blood pool in the place of the resected tumor (red arrows in the figure), which is not found in template image;

2. around the tumor resection areas, there is usually some recurred tumor (blue arrows in the figure), which is also not seen in template image.

**Result**  The mutual-saliency component in DRAMMS helps alleviate this problem – DRAMMS automatically finds those regions that have less chance to find correspondences cross images, and use them with less weight in registration. For example, Fig. 86(d) shows mutual-saliency has effectively assign low weights in those regions in target image space that correspond to outlier regions in source image space (Fig. 86(a)). The nice thing is that, this mutual-saliency calculation is fully automatic, doesn’t need prior knowledge,
or pre-segmentation of tumor or other outlier regions, and also doesn’t need any human intervention or initialization.

Fig. 87 shows the same set of results in coronal view.

Command

```
dramms -S src_Cardiac1.img -T trg_Cardiac2.img
   -O src2trg.hdr   -D def_S2T.nii.gz
   -g 0.4   -c 2
```

**Resources Needed**  Registering this pair of 3D images (target image 256x256x181 voxels, 1.0x1.0x1.0 mm⁻³/voxel) takes 8.8 GB memory and finishes in 33 minutes in Linux OS (2.80GHz CPU).

If one can afford less memory, please use `-u` option to choose memory usage in different levels (the lowest being about 1/4 of maximum memory used). This may however slightly reduce registration accuracy.

**Note**  In this example, we only presented registration of images with tumor recurrence. We have tried to use DRAMMS for registering original tumor images. Sometimes succeeded sometimes failed. The main difference is that, original images usually has bigger and more tumors regions, and more mass effect, edema infiltration that further compound the problem. So it seems fair to say that DRAMMS is good at registering images with lesions, acceptable for images with small number, small size and less infiltration tumors, but may fail for images with large number, large size and large mass effect tumors.

**A.5.2.6. Tutorial 8: Longitudinal Mouse Brain MR Images**

**Introduction**  Due to the its similarity to human brain and its structural simplicity, we often study early-stage mouse brain development to help better infer human brain develop-
Figure 86: Registration of a brain image with tumor recurrence to a normal brain template. Without segmentation/initialization/prior-knowledge, the automatically-calculated mutual-saliency map (d), defined in target image space, has effectively assigned low weights to those regions that correspond to those outlier regions (pointed out by arrows) in source image (a). This way, the negative impact of outlier regions is largely reduced; registration is mainly guided by regions that can establish good correspondences.
Figure 87: The same set of results as in Fig. 86 but in coronal view.

a) Source image (original tumor resected, new recurred)

b) Target image (a normal template)

c) Registered image

d) Mutual-Saliency Map (defined in target image space)
ment. Registration is thus needed to normalize mouse brain images at different time points. Challenges are large structural change or even development of new structures. In the example below, the earlier brain has not yet developed sophisticated cerebellum structures than the other relatively older brain.

![Registration of longitudinal mouse brain images to follow mouse brain development.](image)

Figure 88: Registration of longitudinal mouse brain images to follow mouse brain development.

**Result**

**Command**

```
dramms -S src_mouseDay2.hdr -T trg_mouseDay10.hdr -O src2trg.nii.gz -D def_S2T.nii.gz -g 0.2 -c 2
```
Other Options  If the default parameter does not give good results, it is usually caused by different structures in different life stages. That is, structures may change greatly in shape and size, and may even be missing in images of earlier stage.

Possible solutions include

1. increase search range in DRAMMS (setting control point spacing via -x, -y and/or -z options);

2. use initialization to de-couple a large deformation into several smaller ones (usually known as “geodesic registration”). In such case, one can use deformation from A to B to better initialize the calculation of deformation from A to C (suppose A,B,C are in time order). DRAMMS can take initial deformation by -d option. That is,

From time A to B:

```
dramms -S timeA.hdr -T timeB.hdr
        -O timeA2B.hdr -D def_timeA2B.nii.gz
        -g 0.2 -c 2
```

From time A to C using A to B as initialization ("-d" option):

```
dramms -S timeA.hdr -T timeC.hdr
        -O timeA2C.hdr -D def_timeA2C.nii.gz
        -g 0.2 -c 2
        -d def_timeA2B.nii.gz
```

Resources needed  Registering this pair of 3D images (target image 300 x 300 x 200 voxels, 0.0625 x 0.0625 x 0.0625 mm^3/voxel) takes 8.6 GB memory and finishes in 43 minutes in Linux OS (2.80GHz CPU).

If one can afford less memory, please use -u option to choose memory usage in different
levels (the lowest being about 1/4 of maximum memory used). This may however slightly reduce registration accuracy.

A.5.3. 3D Cardiac Images

A.5.3.1. Tutorial 9: Cardiac MRI (Cross-Subject, Heart Regions only)

Introduction  Cardiac image registration across subjects can help us normal a population of hearts into a common heart template. It applies to cardiac atlas construction, atlas-based heart labeling. It is also one of the first steps for differentiating sub-populations (like normal versus abnormal).

![Registration of a pair of 3D cardiac MR images](image)

**Figure 89**: Registration of a pair of 3D cardiac MR images (different subjects, short-axis, end-diastole). The non-heart structures have already been removed, making registration much easier than keeping all the non-heart structures.

Result

Command

```
dramms -S src_Cardiac1.img -T trg_Cardiac2.img
    -0 src2trg.hdr -D def_S2T.nii.gz
```
**Resources Needed**  Registering this pair of 3D images (target image 120 x 120 x 12 voxels, 1.25 x 1.25 x 8.0 mm^3/voxel) takes 0.6 GB memory and finishes in 2.3 minutes in Linux OS (2.80GHz CPU).

If one can afford less memory, please use `-u` option to choose memory usage in different levels (the lowest being about 1/4 of maximum memory used). This may however slightly reduce registration accuracy.

**Note**  In this example, we have removed all non-heart structures to make registration feasible. When those non-heart structures are kept, especially when different images having different fields-of-view (FOVs) to include different regions of non-heart structures, registration across subjects will be much more difficult. In those cases, DRAMMS, like many other registration methods, is expected to perform poorly.

### A.5.3.2. Tutorial 10: Longitudinal Cardiac MRI

**Introduction**  Registering longitudinal cardiac images of the same subject can help us understand heart motion, and also help propagate heart labeling over time.

The difficulty of registering longitudinal cardiac images of the same subject is in-between the difficulty in registering pure hearts and the difficulty in registering all structures (including non-heart structures) across subjects. Usually, the default setting of dramms will do a reasonably good job.

![Figure 90: Registration of longitudinal cardiac images of the same subject.](image)
Result

Command

dramms -S src_CardiacTP3.hdr -T trg_CardiacTP10.hdr
       -O src2trg.nii.gz -D def_S2T.nii.gz

Resources Needed  Registering this pair of 3D images (target image 120 x 120 x 12
voxels, 1.25 x 1.25 x 8.0 mm^3/voxel) takes 0.95 GB memory and finishes in 2.4 minutes in
Linux OS (2.80GHz CPU).

If one can afford less memory, please use -u option to choose memory usage in different
levels (the lowest being about 1/4 of maximum memory used). This may however slightly
reduce registration accuracy.

A.5.4. 3D Breast Images

A.5.4.1. Tutorial 11: Longitudinal Breast Cancer MR Images

Introduction  Registration is often needed to track breast cancer change over time. This
is especially important to monitor breast cancer patient, and track their response to chemotherapy. It is also one of the first steps towards differentiation between responders (subjects who respond to chemotherapy well, whose tumors usually shrink or even vanish) and non-responders.

Below is an example. DRAMMS registration recovers the deformation from post to pre
chemotherapy in this breast cancer patient.

Result  A typical set of longitudinal registration results have already been shown in 54 in
Chapter 4  Fig. 91 below is the repetition of that figure.
Figure 91: Registration of breast cancer images for the same subject, to monitor the effect of chemotherapy in altering breast cancer over time.

a) Source image (post chemotherapy)  
b) Target image (pre chemotherapy)  
c) Registered image (post2pre)
Command

dramms -S src_breastPost.hdr -T trg_breastPre.hdr
    -O src2trg.nii.gz -D def_S2T.nii.gz -g 0.3

Resources Needed  Registering this pair of 3D images (target image 256 x 256 x 64 voxels, 0.78 x 0.78 x 2.30 mm³/voxel) takes 9.1 GB memory and finishes in 81 minutes in Linux OS (2.80GHz CPU).

If one can afford less memory, please use \(-u\) option to choose memory usage in different levels (the lowest being about 1/4 of maximum memory used). This may however slightly reduce registration accuracy.

A.5.5. 3D Prostate Images

A.5.5.1. Tutorial 12: Histological and MR Images of The Same Prostate

Introduction  Histological image, because of its \textit{ex vivo} and \textit{microscopic} nature, usually provides structural or pathological ground-truth. Finding MR regions corresponding to histological ground-truth regions will help us better understand how a certain structure or pathology appears in MRI. This is often the building block for future labeling or diagnosing new subjects from their MR images.

Histology and MRI registration is in general a very challenging multi-modality registration task. Difficulties are:

1. completely imaging protocols, intensity distribution and image contrasts. This often leads to loss of consistent relationship between intensity distributions of two images, which violates the underlying assumption of mutual-information based classic multi-modality registration approach.
2. structural change due to loss of water (dehydration) in ex vivo histology preparation and section.

3. histological images are sectioned and scanned slice by slice, and stack to together. This sometimes known as 2.5D image has partially lost 3D integrity/continuity as seen in 3D MR images.

4. partial loss of correspondences (histological cuts) due to sectioning in producing histological images.

Below we show an example of using DRAMMS to somehow alleviate some of those issues, mostly because of attribute matching design and mutual-saliency weighting components.

![Example Images](image_url)

**Figure 92**: Registration of multi-modality images of a same prostate. Dashed circles outline the corresponding structures.

**Result**

**Command**

```
dramms -S src_histology.hdr -T trg_MRI.hdr
     -O src2trg.nii.gz -D def_S2T.nii.gz
     -g 0.25 -c 2
```

**Other Options** Despite this success example, registration of histological and MR images is a quite challenging task that even affine registration may often fail. Because of that,
DRAMMS may fail too.

One solution is to carefully re-do affine registration outside DRAMMS scope. After having obtained a reasonable affine registration result, one can input the affinely registered images (src2trg\_affine.nii.gz and trg.nii.gz) into DRAMMS, meanwhile turning off the affine part within DRAMMS by the \(-a \ 0\) option.

**Resources Needed**  
Registering this pair of 3D images (target image 256 x 256 x 64 voxels, 0.16 x 0.16 x 0.4 mm\(^3\)/voxel) takes 6.7 GB memory and finishes in 39 minutes in Linux OS (2.80GHz CPU).

If one can afford less memory, please use \(-u\) option to choose memory usage in different levels (the lowest being about 1/4 of maximum memory used). This may however slightly reduce registration accuracy.

### A.6. Frequently Asked Questions

**A.6.1. Which parameters shall I tune first to get different registration results?**

If you find deformation or registered results too smooth or too aggressive (curly), you can set lower or higher smoothness weight through \(-g\) option of `dramms`. This value is usually between 0 (super aggressive) and 1 (super smooth), with default being 0.2.

If you find deformation not enough to cover large structural variations (may be seen in registering brains with skull, or registering longitudinal images of large structure change), you can enlarge search range by setting bigger values...
of control point spacing via -x, -y and -z options.

Some images are so different that even affine registration may fail. Typical examples include histology-to-MRI registration, prostate/cardiac image registration of different subjects when all surrounding non-prostate/non-cardiac structures are present and FOVs are different. In those situations, failure in affine part within DRAMMS will spread to the subsequent deformable part of DRAMMS. So our suggestion is to carefully re-do affine registration separately outside DRAMMS. Until we get a reasonable affine registration, we can input the affinely registered images (src2trg affine.nii.gz and trg.nii.gz) into DRAMMS, and skip the affine part within DRAMMS (by setting -a 0 option).

Please use dramms --help in command window to display a complete list of parameters available for tuning.

A.6.2. What if I want to re-direct or save all intermediate results?

During runtime, dramms will generate and temporarily save all intermediate results into an unique and automatically-generated sub-directory under the directory where dramms is launched. Those intermediate results include Gabor attribute images, results before and after registration in each image resolutions.

If you want those run-time intermediate results to be re-directed (temporarily saved) to another directory other than the directory where dramms is launched, you can use -I <your_preferred_dir> at the end of dramms main command.

No matter where they are temporarily saved, those run-time intermediate results will be automatically deleted in default settings after dramms finishes. If you want rather keep them for further analysis, you can do so by adding -i option at the end of the dramms command.
A.6.3. How can I reserve memory for dramms command?

If you are running `dramms` in single computer, you don’t need to reserve memory. The `dramms` command will have access to up to the maximum memory in your computer.

If you are submitting `dramms` command via SGE to high-performance-computing servers, you can reserve memory via adding `-l h_vmem=NG` option to `qsub` (just replace N with the number of GB you want to reserve).

A.6.4. How can I reduce memory use, and/or running time?

The `-u` option of `dramms` allows users to use different levels of memory (1—1/4 of the maximum memory used).

To reduce running time, one can use less iterations (`-k` option, default 10), less discrete samples during optimization (`-n` option, default 5), and use higher weights for deformation smoothness (set `-g` option greater than 0.5).

Please note, using less memory and less running time may lead to slight decrease in registration accuracy.

A.6.5. Can DRAMMS do 2D-to-3D/4D/group-wise registration?

No. At this version DRAMMS only applies to 2D-to-2D and 3D-to-3D pair-wise registrations.
A.6.6. Can I develop new algorithms based on DRAMMS framework?

Yes. DRAMMS is open source and is a general framework where each of its building components can be easily changed.

In current implementation, we describe each voxel by Gabor texture attributes. We match voxels by minimizing their squared differences in Gabor attributes. Further, we modulate this matching process by mutual-saliency weighting at each voxel pair based on their matching reliability.

Those three major components can be easily changed in this framework. For example, developers can try different attribute descriptors other than Gabor features; can try different voxel (dis)similarity definition other than squared different of attributes; and can try different voxel-wise weighting (or even equal weighting) other than mutual-saliency weighting. To change them, developers need to change `CalculateGaborTextures.cxx` and `Deform3D.cxx` programs under `src/tools/` directory in the downloaded software package.
A.7. Publications Related to the DRAMMS Software

A.7.1. Methodology

Ou et al. (2011)

(Ou et al., 2009b)

A.7.2. Validations

Cardiac Registration

(Ou et al., 2012)

A.7.3. Applications

Longitudinal Cardiac Segmentation

(Bernardis et al., 2012)

Brain MRI Skull Stripping

(Doshi et al., 2012)

NeuroDegenerative Disease Study

(Zanetti et al., 2012)

Extraction of Landmark Correspondences

(Sotiras et al., 2010)

(Ou et al., 2010)
A.8. People Contributing to the DRAMMS Software

A.8.1. Advisor

- Christos Davatzikos

A.8.2. Software Development

- Yangming Ou
- Andreas Schuh
- Aristeidis Sotiras

A.8.3. Algorithm Development

- Yangming Ou
- Aristeidis Sotiras
- Nikos Paragios
- Christos Davatzikos
A.8.4. Acknowledgement to Libraries

- FSL flirt tool for affine registration (Analysis Group, FMRIB, Oxford, UK; FSL license)
- NiftiClib (Bob Cox, Rick Reynolds; NiftiClib license)
- FastPD Discrete optimization (Nikos Komodakis, Nikos Paragios; FastPD license)
- Feature extraction, data structure (Yiqiang Zhan, Dinggang Shen)

A.8.5. Software Testers

Thank you for your precious feedback and suggestions during the development of this software!

Within SBIA

- Hamed Akbari
- Harsha Battapady
- Vanessa Clark
- Xiao Da
- Jimit Doshi
- Harini Eavani
- Guray Erus
- Bilwaj Gaonkar
• Madhura Ingalhalikar

• Dongjin Kwon

• Drew Parker

• Alex Smith

• Erdem Varol

• Steffen Wachenfeld

• Dong Hye Ye

Outside SBIA

• Nikolaos Koutsouleris (Ludwig-Maximilians-University, Germany)

• Helene Langet (GE Healthcare/Ecole Centrale de Paris, France)

• Lena Rademacher (The University Hospital Aachen, Germany)

• Marcus Zanetti (University of San Paulo, Brazil)


B. Fischer and J. Modersitzki. A unified approach to fast image registration and a new


C. Jongen, J. van der Grond, and L. Kappelle. Utrecht diabetic encephalopathy study


285


