Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2006-005 January 25, 2006

A Unified Information Theoretic Framework for Pair- and Group-wise Registration of Medical Images

Lilla Zollei

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

A Unified Information Theoretic Framework for Pair- and Group-wise Registration of Medical Images

by

Lilla Zöllei

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

January 2006

Author .....
Department of Electrical Engineering and Computer Science
January 05, 2006

Certified by .....
Eric Grimson
Bernard Gordon Professor of Medical Engineering, CSAIL, MIT
Thesis Supervisor

Certified by .....
William Wells
Associate Professor of Radiology, Harvard Medical School
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

A Unified Information Theoretic Framework for Pair- and Group-wise Registration of Medical Images

by
Lilla Zöllei

Submitted to the Department of Electrical Engineering and Computer Science
on January 05, 2006, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science and Engineering

Abstract

The field of medical image analysis has been rapidly growing for the past two decades. Besides a significant growth in computational power, scanner performance, and storage facilities, this acceleration is partially due to an unprecedented increase in the amount of data sets accessible for researchers. Medical experts traditionally rely on manual comparisons of images, but the abundance of information now available makes this task increasingly difficult. Such a challenge prompts for more automation in processing the images.

In order to carry out any sort of comparison among multiple medical images, one frequently needs to identify the proper correspondence between them. This step allows us to follow the changes that happen to anatomy throughout a time interval, to identify differences between individuals, or to acquire complementary information from different data modalities. Registration achieves such a correspondence. In this dissertation we focus on the unified analysis and characterization of statistical registration approaches.

We formulate and interpret a select group of pair-wise registration methods in the context of a unified statistical and information theoretic framework. This clarifies the implicit assumptions of each method and yields a better understanding of their relative strengths and weaknesses. This guides us to a new registration algorithm that incorporates the advantages of the previously described methods. Next we extend the unified formulation with analysis of the group-wise registration algorithms that align a population as opposed to pairs of data sets. Finally, we present our group-wise registration framework, stochastic congealing. The algorithm runs in a simultaneous fashion, with every member of the population approaching the central tendency of the collection at the same time. It eliminates the need for selecting a particular reference frame *a priori*, resulting in a non-biased estimate of a digital template. Our algorithm adopts an information theoretic objective function which is optimized via a gradient-based stochastic approximation process embedded in a multi-resolution setting. We demonstrate the accuracy and performance characteristics of stochastic congealing via experiments on both synthetic and real images.

Thesis Supervisor: Eric Grimson
Title: Bernard Gordon Professor of Medical Engineering, CSAIL, MIT

Thesis Supervisor: William Wells
Title: Associate Professor of Radiology, Harvard Medical School

Thesis Committee: John Fisher
Title: Principal Research Scientist, CSAIL, MIT

Thesis Committee: Alan Willsky
Title: Professor of Electrical Engineering, LIDS, MIT

Acknowledgments

The acknowledgement section: as always, the most difficult but undoubtedly the most heartwarming section to write. It fills me with great happiness to think about all the six years that I spent at MIT and all the help, supervision, guidance and friendships that I have found there. Even though the beginning was quite intimidating, I dearly enjoyed my graduate life in Cambridge.

With respect to my dissertation work, first and foremost, I would like to acknowledge the guidance of my committee members. I truly appreciate the timely feedback that I received from Prof Willsky, the caring supervision and protection from Prof Grimson as a supervisor over all my years at MIT and all the thought-provoking brainstorming sessions that I had with John Fisher (wherever we stumbled into a white-board). My special thanks go to Sandy Wells, who co-supervised both my Masters and my PhD studies at the lab. His unfailing support, care and insights helped me through the hardest times of research and (post-graduate) decision making. And not only did I truly enjoy working with him for all these years in the lab, but we also played in many ice hockey games together which remain dear memories, too.

Among my close research collaborators, I would like to mention Erik Learned-Miller, Andrea Mewes, Neil Weisenfeld, Simon Warfield and Bruce Fischl. Their research expertise and enthusiasm greatly inspired me and our joint projects further enhanced my knowledge about numerous theoretical and medical topics. In general, I am very honored having had the chance to participate in several research initiatives at the SPL, a unique and incredibly exciting lab headed by Dr. Ron Kikinis and Dr. Ferenc Jolesz.

I owe a tremendous amount of gratitude to my unofficial thesis readers: Eric Cosman, Corey Cemper, Lauren O'Donnell, István Zöllei, Kilian Pohl, Jérôme Huber, Sonya Chu, Florent Ségonne and Wanmei Ou. Given all the time pressure that I had before finishing my dissertation, their input and constructive comments were invaluable. Eric, Lauren and Corey deserve an extra load of appreciation: Eric and Lauren for completely tearing apart some of my chapters demanding a more precise and clearer explanation, and Corey for covering the largest number of pages with high precision in very little time.

Besides working on the final document, I got a great amount of support from my friends and the members of the Vision Group who made the every day life in Cambridge so much more fun. I would like to specifically mention my three *best office-mates*: Florent Ségonne, Eric Cosman and Anna Custo. Although Stata D430 was quite a tiny office for all of us, and our musical preferences are highly non-overlapping, we managed to survive our time together with lots of laughter, gossips and delicious dinners. Even with our dear rodent guests, I will think of that office and all the fun activities that we did together very fondly. One more office-mate deserves to be added here: Dave Gering. He was also a lot of fun to have around, before he decided to move back to Wisconsin. In the lab, outside the small world of my office, I was extremely lucky to befriend a set of other great pals: Kilian Pohl, Eduardo Torres-Jara and Vikas Sharma. Their humor, optimism and unconditional

support have all contributed to endless happy moments.

My life in Boston would not have been complete without the small but lively Hungarian community. Besides my official involvement in the life of the student community, I truly cherished the time that I spent with a smaller group of friends: Horváth Adrián, Mikó Márta, Ilkey Csilla, Cserny Réka, Kovács Tamás, Szepessy Edit, Nemoda Zsófia, Falus Péter, Farkas Dóra and Ittzés Laura. With them I could always sing silly Hungarian songs when in a nostalgic mood, cook some delicious dinners, do some Hungarian folk dancing, watch Harvard basketball and waterpolo games. I am greatly delighted that I met many of you right after my arrival at MIT and I hope to be in touch with you all for a very long time.

I found another great group of friends as I stumbled upon the basketball team of the European Club. Our practices, games and late FZB parties provided me with lots of great memories and friendships that I value sincerely. We have had many players and lots of fun, but I would like to personally mention the ones that were closest to me: Birgit Schoeberl, Hélène Karcher, Jan Lammerding and David Garcia Alvarez.

Through Birgit I also had the opportunity to meet the wonderful Spring Street gang, including Franciska Leite, Tiago Castro Riberio and Florian Altmann. With them I share many skiing / hiking trip memories and late night party stories.

Two more couples deserve my warmest appreciation for their unfailing care, love and support throughout my graduate studies. They are: Katalin Révész and Zsolt Belánszky, my closest confidants from home; and Shannon Dawson and Clemens Hrouda, my closest *grown-up* friends in the US.

Je voudrais aussi mentionner l'amitié et l'amour de Jérôme Huber qui j'ai rencontré il y a quatre ans à MIT. Jérôme, merci beaucoup pour toute la confiance et l'encouragement que j'ai recu de toi pendant toutes ces années!

Last, but unquestionably not least, I would like to acknowledge the eternal support and love that I received from my family. From so many miles away, their encouraging words never failed to reach me and their readiness to help and console was a rock-solid pillar throughout all these years. "Anyu, Apu és Pisti! Végtelenül köszönöm minden biztatásotokat és bizalmatokat amit belém fektettetek ez a hosszú és néha igen embertpróbáló időszak alatt. Szeretetetek és gondoskodásotok nap mint nap újabb és újabb erőt adott a gondok legyőzéséhez és még mosolygósabbá tett a boldog pillanatokban."

This thesis is dedicated to my late grandmothers, "Anna" and "Irén". Neither of them were able to witness the finishing years of this Transatlantic project, but they were always supportive with respect to all my aspirations. "Nagyikák, köszönök mindent!"

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Medical Image Processing

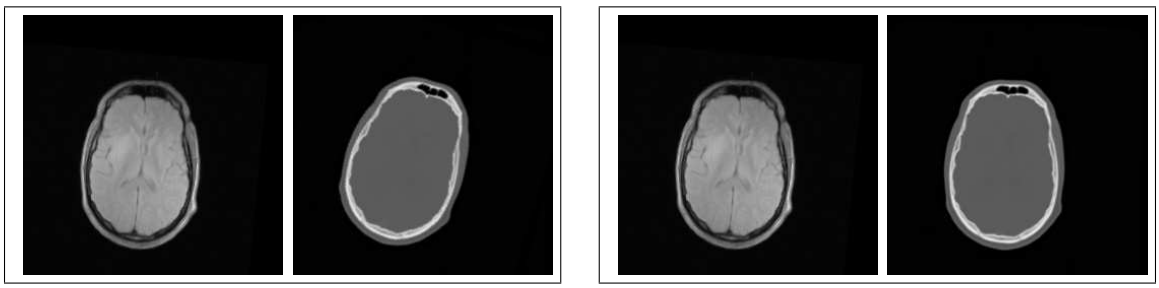
The field of medical image analysis, which includes the post-processing and interpretation of the great variety of image acquisitions that are used for diagnostic and interventional purposes, has been growing rapidly for the past two decades. This acceleration is partially due to a significant growth in computational power, scanner performance, and storage facilities. Given the technological advances, it is possible to process large size data sets with improved speed and accuracy; and it has also become a standard to digitally store most of the acquisitions. These together facilitate data collection initiatives over longer periods of time, potentially allowing for efficient analysis of groups of data sets as opposed to just individual scans.

Advancements in medical image analysis can also be explained by an unprecedented increase in the amount of data sets accessible for researchers. As the number of available imaging modalities is rapidly growing, a more detailed and complete characterization of the imaged anatomies is becoming attainable. Consequently it is not only the structural features of the examined anatomies that are thoroughly described, but information about their functional properties can also be acquired. Besides the analysis of multiple acquisitions of the same subject, comparisons of images across subjects or groups of subjects are also becoming popular.

Medical experts traditionally rely on manual comparisons of images, but the abundance of information now available makes this task increasingly difficult. Such a challenge prompts for more automatic processing of the images, which could produce a summary and potentially a visual display of all the available information relevant to the examined body parts.

In order to carry out any sort of comparison between multiple medical images, one needs to identify the proper correspondence between them. A processing step of this sort then allows us to follow the changes that happen to anatomy throughout a particular time interval, to compare differences between individuals, or to acquire

complementary information from different image modalities that describe the same anatomy. One procedure that achieves such a correspondence, or in other words that allows for such comparisons across input images, is called *registration*. In this dissertation we focus mostly on the analysis and characterization of the different registration approaches. We provide a simple 2D example on Fig. 1-1 and a detailed characterization of the registration task and the state-of-the art in Chap.2.



(a) Unaligned medical images (b) Aligned medical images

Figure 1-1: 2D example of a pair of (a) misaligned and (b) aligned medical images. The images are MRI and CT acquisitions of the brain.

Another key processing method in medical image analysis is called *segmentation*. This procedure is responsible for assigning descriptive labels to all the voxels (3-D pixels) of the input. Such labels frequently correspond to anatomical features or tissue types. Thus, for example, one can imagine a segmentation scenario where the relevant image components are all classified into anatomical categories, such as: white matter, gray matter and cerebro-spinal fluid. In Fig. 1-2, we provide a 2D segmentation example derived from an MRI scan of a pre-term baby brain.

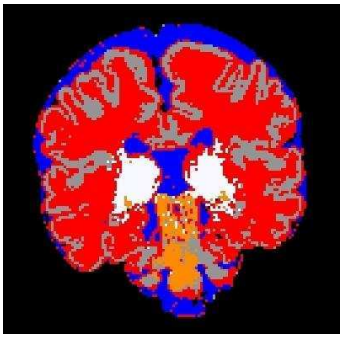


Figure 1-2: Medical image segmentation example: a 2D coronal slice of a baby brain acquisition is segmented into cortical gray-matter (gray), unmyelinated white-matter (red), cerebro-spinal fluid (blue), myelinated white-matter (orange) and basal ganglia (white).

Registration and segmentation processes often play complementary roles in image processing. Informally, when images are aligned and the segmentation of one of them

is known, it is easier to segment the other image. Also, if two images are segmented using the same set of classification labels, the set of corresponding features can aid in identifying the transformation that would put them into alignment. It is also possible to validate the accuracy of registration when the correct segmentation is known and vice versa. In our experimental work, we demonstrate examples of both of these scenarios. We both validate our registration results by examining the resulting segmentation quality and also improve on the quality of image segmentations by aligning them to previously examined data sets.

1.2 Medical Image Modalities

Although most of our analysis and registration algorithms could be applied to a wide variety of image types, in this dissertation, we focus only on medical data sets. In order to avoid confusion and to allow the reader to have a sufficient amount of knowledge about their characteristics, herein we provide a brief overview of the most commonly mentioned volumetric medical image modalities. The information in this section is largely derived from overview publications and books [28, 8, 40, 87].

1.2.1 Magnetic Resonance Imaging

The most predominantly mentioned data type in our current work is Magnetic Resonance Imaging (MRI). This modality was introduced roughly 30 years ago and since then its image quality and reliability have been improving. The role of MRI both in research and clinical practice is significant, and in many western countries, it is becoming a commonly used form of medical imaging. A more widespread use of this modality is prevented by the fact that the acquisitions at the moment are very expensive, approximately $1000 per scan. In clinical practice, MRI is used to distinguish pathologic tissue (such as a brain tumor) from normal tissue. One of the practical advantages of an MRI scan is that, according to current medical knowledge, it is harmless to the patient. It utilizes strong magnetic fields and non-ionizing radiation in the radio frequency range. Although it is considered to be a safe imaging modality, proper precautions must be taken in order to eliminate the presence of ferromagnetic materials in the vicinity of the magnet of the imaging device, otherwise, fatal accidents can occur [7].

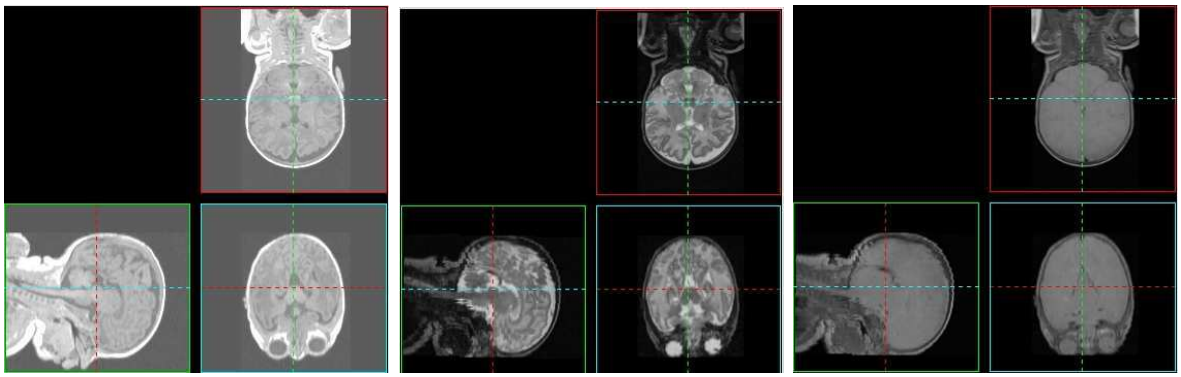
With conventional MRI systems, most acquisitions take a minimum of 2 minutes, but with some advanced technology, sub-second scans can also be obtained. Magnetic resonance images are also associated with high contrast resolution (the ability to distinguish the differences between two similar but not identical tissues) and their spatial resolution is comparable (the ability to distinguish two structures an arbitrarily small distance from each other as separate) with that of Computed Tomography images. The latter image modality is introduced in Sec.1.2.2.

Magnetic resonance imaging is based on the principles of nuclear magnetic resonance (NMR). The term *nuclear* was dropped from the name of the imaging technique because of the negative connotations associated with the word nuclear in the late 1970's. Over the years, MRI has also advanced from being a tomographic imaging technique to a volume imaging method.

MRI imaging most frequently relies on the relaxation properties of excited hydrogen nuclei in water. When the object to be imaged is placed in a powerful uniform magnetic field, the spins of the atomic nuclei with non-zero spin numbers within the tissue all align either parallel or anti-parallel to the magnetic field. The magnetic dipole moment of the nuclei then precess around the main magnetic field. The frequency with which the dipole moments precess is called the Larmor frequency. The tissue is then briefly exposed to a sequence of pulses of electromagnetic energy (RF pulse) in a transverse plane that is perpendicular to the magnetic field. This causes some of the magnetically aligned hydrogen nuclei to assume a temporary non-aligned high-energy state. The frequency of the pulses is governed by the Larmor Equation describing the relationship between the angular frequency of a precessing proton, and the strength of the magnetic field.

The following describes one way of acquiring images with MRI. In order to selectively image the different voxels of the material in question, three orthogonal magnetic gradients are applied. The first is the *slice selection*, which is applied during the excitatory RF pulse. Next comes the *phase encoding* gradient, and finally the *frequency encoding* gradient, during which the tissue is imaged. Most of the time, the three gradients are applied in the X, Y, and Z directions of the machine; however, MRI is especially useful because various combinations of the gradients can be combined during the process so that slices can be taken in any orientation.

As the excited nuclei relax and realign, they emit energy which provides information about their environment. The realignment with the magnetic field is termed longitudinal relaxation and the time required for a certain percentage of the tissue nuclei to realign is termed T1. This is the basis of *T1-weighted* imaging. *T2-weighted* imaging relies upon local dephasing of spins following the application of the transverse energy pulse; the transverse relaxation time is termed T2. Both T1- and T2-weighted images are frequently acquired for most medical examinations. Often, a paramagnetic contrast agent, a gadolinium compound, is administered, and both pre-contrast T1-weighted images and post-contrast T1-weighted images are obtained. Another type of MRI, the *PD-weighted* (or proton density-weighted) image is produced by controlling the selection of scan parameters to minimize the effects of T1 and T2. This results in an image dependent primarily on the density of protons in the imaging volume. Proton density contrast is a quantitative summary of the number of protons per unit tissue. The higher the number of protons in a given unit of tissue, the brighter the signal in the proton density contrast image. An example of each of these modalities is displayed in Figure 1-3. The acquisitions are aligned and were taken of the same subject.



(a) T1-weighted MRI (b) T2-weighted MRI (c) PD-weighted MRI

Figure 1-3: Three orthogonal views of three different types of MRI: (a) T1-weighted, (b) T2-weighted, and (c) PD-weighted. The data volumes are aligned and are taken of the same subject.

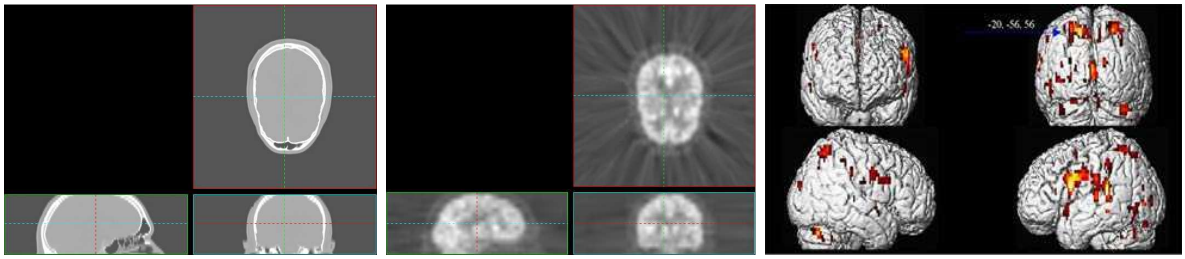
Many other specialized types of MRI exist, including diffusion weighted MRI and magnetic resonance angiography. As they are not directly relevant in our discussion, we do not describe them in detail here.

For a more detailed description of the physics and the intensity characteristics of this modality, we refer the reader to three excellent sources [28, 40, 87].

1.2.2 Computed Tomography

Another significant data modality that we use in our registration examples is Computed Tomography (CT). Since its introduction in the 1970s, CT has become an important tool in medical imaging to supplement projection X-rays and medical ultrasonography. Although it is still quite expensive (approximately $200 per scan), it costs significantly less than the MRI acquisitions. CT has become the gold standard in the diagnosis of a large number of different diseases. It is used in medicine as a diagnostic tool and as a guide for interventional procedures. Sometimes contrast materials such as intravenous iodinated contrast are used. This is useful to highlight structures such as blood vessels that otherwise would be difficult to distinguish from their surroundings. Using contrast material can also help to obtain functional information about tissues.

Computed tomography was originally known as computed axial tomography (CAT). After a series of two-dimensional X-ray images are taken around an axis of rotation, digital processing is used to generate a three-dimensional image of the internal structures of the examined object. As a result, the scan time has decreased, and the ability to reconstruct images (for example, to look at the same location from a different angle) has increased over time. Images that used to take hours to acquire and days to process are now accomplished in seconds, and the number of cross sectional images



(a) CT (b) PET (c) fMRI

Figure 1-4: Three different medical image modalities: (a) CT, (b) PET, and (c) fMRI activation map. (The fMRI activation image was acquired from [www2.uibk.ac.at]). Three orthogonal views are displayed for the CT and the PET intra-subject data volumes. The fMRI acquisitions are not related to the same subject; the activation maps are projected onto the reconstructed cortical surface.

that can be produced has increased from about a dozen to many hundreds. Still, the radiation dose from CT scans is several times higher than conventional X-ray scans.

For further details on tomographic reconstruction techniques and the improvement in CT imaging technology, we refer the reader to [8, 40, 87].

1.2.3 Functional Imaging

Finally, we briefly mention two other data modalities that occur throughout our discussions: Positron Emission Tomography (PET) and Functional Magnetic Resonance Imaging (fMRI). In contrast to the MRI and CT data sets that provide a detailed representation of the structural characteristics of the anatomy, these two modalities are used for the analysis of functional properties of the imaged anatomical areas. In the brain, for example, hemodynamic phenomena are presumed to be related to neural activity. Thus PET and fMRI techniques can be used to determine what the brain is doing when subjects perform specific tasks or are exposed to specific stimuli. These images are increasingly read alongside the structural acquisitions, as the combination of the two provides both anatomic and functional information about the examined organs.

In PET imaging, following the injection of a short-lived radioactive tracer isotope into the living subject, changes in regional blood flow in various anatomic structures can be quantified by indirectly observing radioactive decay. The resulting map shows the tissues in which the injected molecular probe has become concentrated. PET is used heavily in clinical oncology (medical imaging of tumors and the search for metastases) and in human brain and heart research, but its use as a technique for scientific investigation is limited because of the danger imposed on the participants by the injection of radioactive material. Although they provide quantitative information about biological processes, the resolution in these images tends to be low, and thus their automatic analysis can be a challenging task. In Figure 1-4 (b), we show a PET

head image.

Functional Magnetic Resonance Imaging is the name of a relatively new method for using MRI to observe the hemodynamic response related to neural activity in the brain or spinal cord of humans or other animals. It is one of the most recently developed forms of brain imaging. fMRI has a better temporal and spatial resolution than PET to examine blood oxygenation levels and blood flow rates. Frequently, after image acquisition, an fMRI activation map is computed from the time series of *echoplanar* MRI images. Such a map then highlights certain anatomical areas that were most probably active during the imaging process. An example of a 3D activation map is displayed in Figure 1-4 (c).

1.3 Problem Statement and Contributions

In this dissertation, our main focus lies in finding anatomical correspondences between a set of observed image acquisitions. We are interested in both the theoretical analysis and the experimental behavior of a select group of alignment strategies. Although there has already been a large number of registration approaches introduced in the medical imaging literature, the advantages of each have been mostly demonstrated individually and there has been only limited effort put into comparing them all in a unified framework. We believe that a detailed analysis of the registration formulations has been missing and we have been motivated to introduce one. As most of the principled registration approaches rely on statistical formulations of the image intensity information, we construct a unified statistical framework. It facilitates a better understanding of the implicit and explicit assumptions the registration methods makes and it allows us to compare their relative strengths and weaknesses.

Based upon the number of images that are to be aligned, we differentiate between pair-wise and group-wise registration tasks. The former problem has been studied for more than a decade while the latter has just recently become a topic of interest. On Fig. 1-1 we present a simple case of pair-wise rigid registration between an MRI and a CT image and on Fig. 1-5 we demonstrate one of our group-wise registration results that is explained in more details in Chapter 5.

In our work, we first carefully analyze the pair-wise registration problem, and then demonstrate how our framework can be naturally extended to the group-wise scenario.

Additionally, we also define new alignment methods. As a result of our unified study of registration algorithms, we identified certain aspects of the currently existing techniques that could be improved upon. In the pair-wise registration scenario, for example, using information from previously registered data sets is shown to increase the robustness and the capture range of the technique. However, this information can also limit the alignment accuracy, unless one is flexible when using such prior constraints. In our new framework, we encode any previous knowledge to the current registration problem by defining a probability distribution (instead of a fixed model) on the currently observed joint statistics.

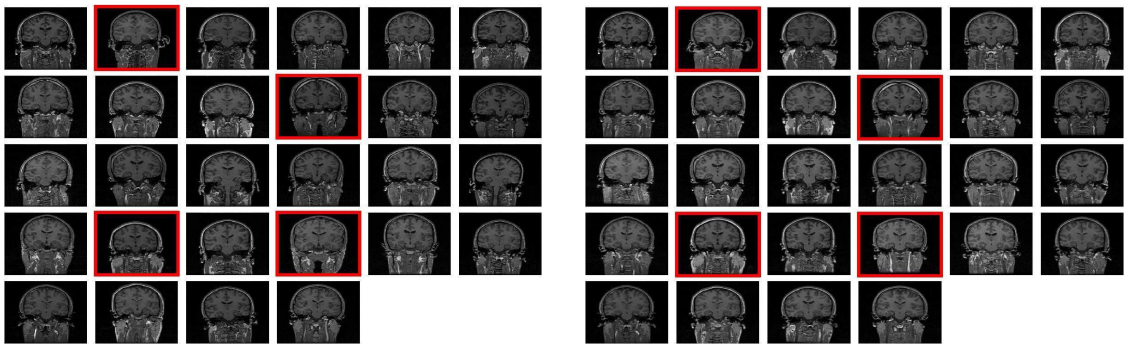


Figure 1-5: An adult brain data set of 28 MRI volumes. Central coronal slices of the input images (a) before and (b) after group-wise registration.

In the case of group-wise registration problems, it is often difficult to capitalize on the results of former high accuracy alignments. Instead, an emphasis is placed on identifying a computationally efficient approach that could be applied even in the case of very large input data sets. We capitalize on a registration criterion whose statistical computations remain one-dimensional even with an increasing size of the input data set. Using that technique, we also manage to address certain aspects of defining digital anatomy atlases.

In summary, our main contributions to the registration field include:

- a unified statistical framework to compare pair-wise registration objective functions;
- a novel pair-wise registration objective function that incorporates information from both previously aligned data sets and the current observations;
- the extension of the unified statistical framework to include group-wise registration analysis; and
- the introduction of a novel group-wise registration framework that is suited for aligning gray-scale medical data sets in a computationally efficient manner.

1.4 Thesis Road Map

The seven chapters of this dissertation are organized in the following manner. Chapter 2 gives an overview of the image registration problem with the state-of-the-art alignment approaches in the medical imaging field that have been proposed to address it. We make an explicit distinction between the pair-wise and the group-wise techniques, and we also introduce digital atlas construction. In Chapter 3, we present a unified information theoretic framework that explains the relationship among a select group of widely used registration methods. Following that analysis, we derive a new

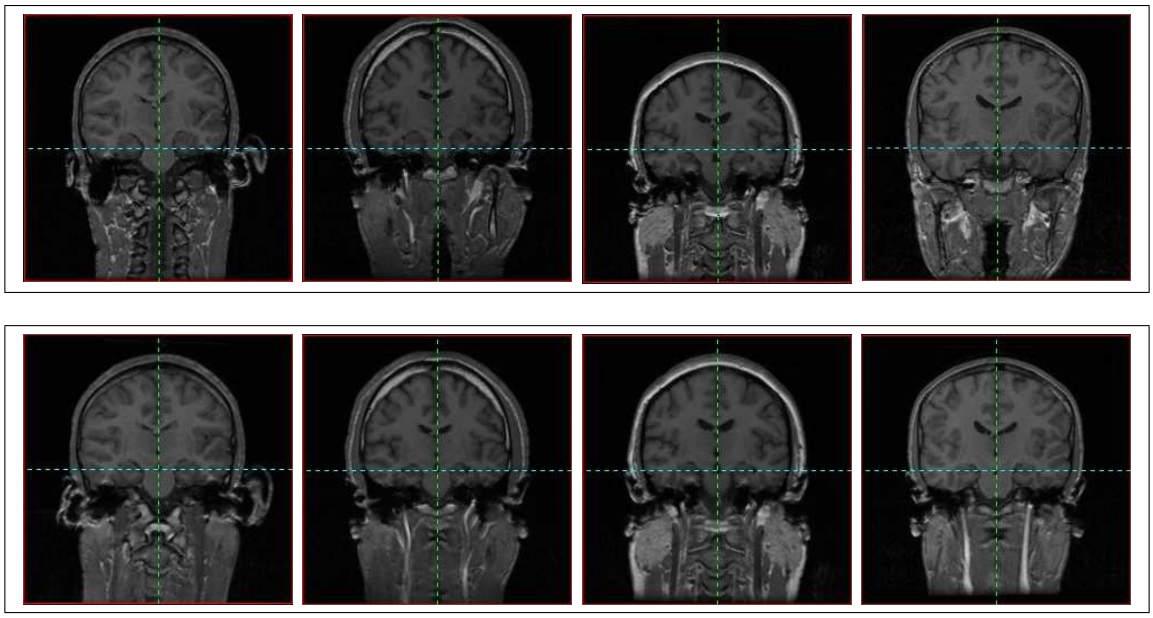


Figure 1-6: Four examples from an adult brain data set of 28 T1-weighted MRI volumes. (The selections are indicated with a red box in Fig. 1-5.) Central coronal slices of the input images were obtained (top row) before and (bottom row) after registration.

registration method that builds on the strength of previously analyzed metrics and present its advantages through a set of probing experiments. In Chapter 4, we generalize the information theoretic framework to group-wise alignment metrics and also present a detailed overview of the registration framework that we selected to apply to the large data set registration problem. In Chapter 5, we present experimental results related to our group-wise registration work. We show its strengths and limitations and carefully validate its results both quantitatively and qualitatively. In Chapter 6, we demonstrate further applications that significantly benefit from our group-wise results. Finally, in Chapter 7, we reiterate the contributions of this thesis and point out future directions that we are interested in investigating closely related to this work.

Chapter 2

Background and Literature Review

In this chapter, we provide a detailed overview of the medical image registration problem. We introduce a useful set of vocabulary terms used in the field and present state-of-the-art approaches in pair- and group-wise registration and the related task called digital anatomy atlas creation. While examining the main ideas and assumptions that have been shaping the rapidly growing number of registration techniques, we also point out key application areas that directly benefit from registration.

2.1 Registration of Medical Data

Registration of medical image data sets presents the problem of identifying a homology or a geometric transformation which maps the coordinate system of one data set to that of another. In other words, it is the problem of finding anatomical correspondence between the input data sets. Such an alignment can facilitate a detailed comparison of the analyzed images, and can also allow for accumulation of information about the same anatomy into a single data coordinate frame. Schematically, we illustrate the registration problem on Fig. 2-1. Given the input data sets to be aligned there are three main aspects of the problem that need to be carefully analyzed. First the appropriate transformation domain needs to be identified; second, a special function (or an objective function or a similarity metric) needs to be identified which evaluates the quality of the current alignment; and thirdly an optimization algorithm needs to be selected in order to efficiently explore the search space. In our work, we focus mostly on the first two of these components. Therefore, after a brief summary of the most common registration problems, we describe their roles in more detail.

2.1.1 A Variety of Registration Scenarios

Perhaps the simplest registration scenario arises when we want to register images acquired of the same anatomy with the same type of imaging device. We identify

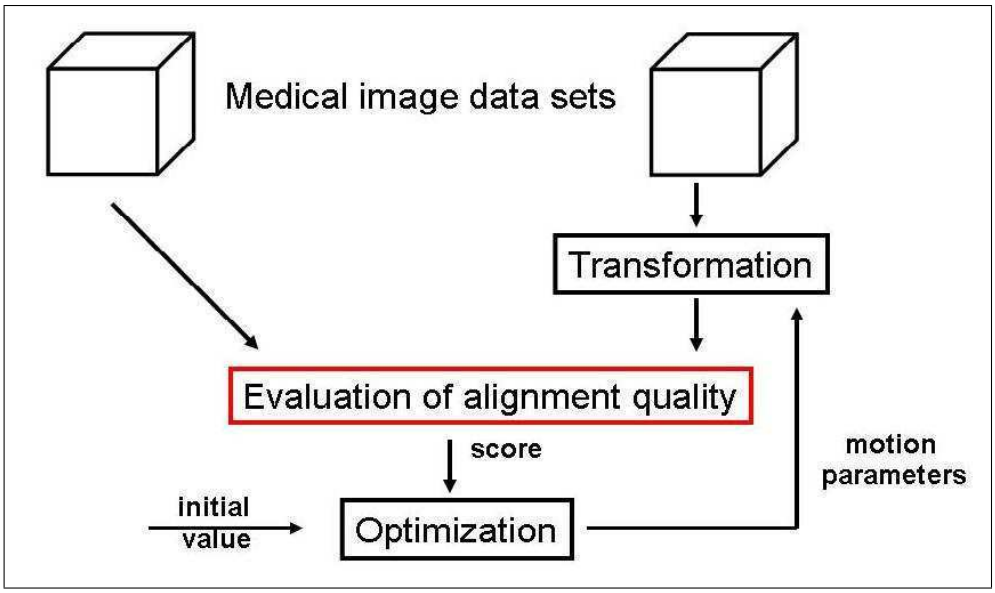


Figure 2-1: A schematic representation of the registration problem with its three main components: transformation; similarity functions; optimization procedure.

such a task as a *uni-modal*, *intra-subject* registration problem. If the subject does not move and the time delay between the two acquisitions is negligible, the alignment is trivial. However, if the image acquisitions are separated by a significant time lag, complex changes could occur to the examined anatomy. For example, the size and/or the shape of the organ might change as a result of disease, or in the case of a clinical data set, certain structures might be missing due to an intervention scheduled between the image acquisitions. In order to detect such changes, or to simply compare the corresponding features, the images need to be properly aligned.

It is also fairly common to image the same anatomy with several different types of imaging device. The different imaging modalities often provide detailed information about non-overlapping properties of the examined anatomy. For example, in the case of anatomical scans, MRI images are able to discriminate with great accuracy between soft tissue types, while CT images are excellent for examining the bony tissues. Besides anatomical features, we can also measure functional properties of certain areas using, for instance, *f*MRI and PET images. Thus, by finding the corresponding areas in various different image representations, we can significantly increase our knowledge about the examined anatomy. This problem, the *multi-modal* intra-subject registration task, requires a more detailed analysis of the observed images as the corresponding anatomical structures might have significantly different intensity profiles, and the intensities that are compared refer to a set of different physical properties.

An even more ambitious task is formulated in the *inter-subject* framework, where we need to compare images of the same anatomy across different subjects. That task requires not only a robust way of identifying common patterns in the input observa-

tions, but also a more complex set of assumptions that could describe the (normal and/or pathological) differences between the inputs. For example, head acquisitions of even just control subjects (those that have not been diagnosed with any particular disease) can demonstrate major differences. When designing a registration method to align them, we need to account for the fact that some structures may not have a precisely corresponding counterpart.

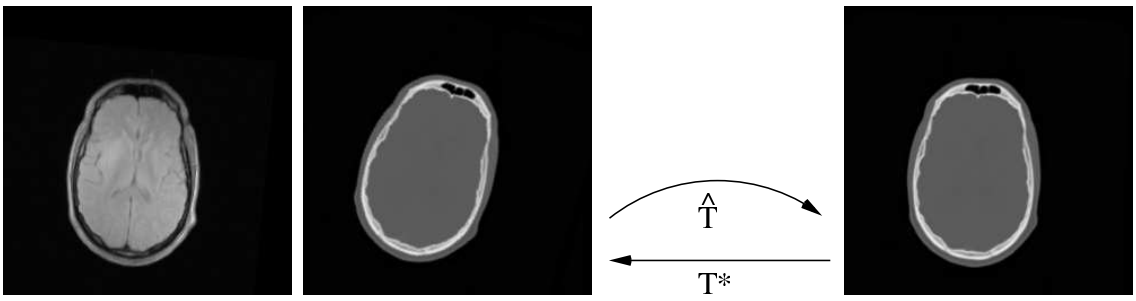
In our theoretical and experimental analysis, we mostly focus on multi-modal intra- and inter-subject registration scenarios.

2.1.2 The Registration Transformations

Based on the nature of the transformations required for alignment, we identify *rigid*, *affine* and *free-form* alignment techniques. In the case of rigid alignment, the motion applied to the input(s) is restricted to rotation and displacement transformations. In the case of affine alignment, scaling and shearing transformations are also permitted. Lastly, in the case of free-form warps, a wide range of relatively unconstrained deformations can be defined, including, for example, a multitude of spline-based methods [57, 56, 44],the demons algorithm [6, 67], and some bio-mechanical deformation models which can facilitate the modeling of tumor growth and the brain shift phenomenon [11]. Consistent linear elasticity models are frequently used to describe more restricted, small deformations [30], while larger deformations can be encoded via approximations to the viscous fluid deformation model [9, 4, 15].

The choice of the appropriate set of transformations is dictated by the nature of the input images and also by the accuracy requirement of the alignment task. Most often, when registering intra-subject images, rigid or affine transformations are sufficient. However, if the time lag between the image acquisitions is rather long or an anomaly and/or an external intervention (e.g. surgery) significantly changes the nature of the imaged anatomy, higher dimensional transformations are preferred. Aligning data volumes of different subjects usually prompts us to use free-form deformations (assuring the recovery of most similarities and differences). Also, rigid transformations are generally sufficient in the case of aligning intra-subject images of bony structures but non-rigid mappings are frequently necessary for soft tissue matching.

Figure 2-2 demonstrates a simple case of a 2D multi-modal registration problem along with the notation we adapt in Chapter 3. The input images are corresponding slices of an MRI and a CT scan of the head that are initially misaligned. We wish to apply a transformation to the CT slice in order to align it with the MRI. More precisely, in mathematical terms, our task is to recover transformation $\tilde{T}$ that best approximates the inverse of T^*, the true unknown transformation responsible for the offset. In this example, it is sufficient to search over the space of rigid transformations as the scans have the same scaling parameter, they were taken of the same subject,



MRI input image CT input image *aligned* CT image

Figure 2-2: A simple 2D example of the multi-modal, intra-subject registration problem. The *observed* input images, an MRI and a CT slice, are not initially aligned. The CT image on the far right has been transformed via $\hat{T}$ and is in proper alignment with the MRI. The unknown transformation that relates the observed CT to the aligned one is T^*. The goal of the registration algorithm is to make $\hat{T}$ be the best estimate of $(T^*)^{-1}$.

Subject 1: T1-weighted MRI Subject 2: T1-weighted MRI

Figure 2-3: Example slices from a 3D uni-modal, inter-subject registration problem. Note that the slices are not corresponding as the image data sets are not currently aligned. The input data sets are T1-weighted MRI images of different subjects. In order to align these images, a non-rigid deformation field needs to be applied.

and no visible external intervention or disease has modified the shape or the main characteristics of the imaged anatomy.

Figure 2-3 presents a more challenging case: a uni-modal, inter-subject registration problem. In order to obtain accurate correspondence, a free-form deformation needs to be used. Not only were the two T1-weighted MRI data volumes acquired of different subjects, but in the case of Subject 2, the development of an easily distinguishable tumor has also modified the shape of the ventricles. Only non-rigid deformations can cope with such differences in the data sets[1].

[1]Our two example images were provided as part of the *Retrospective Image Registration Evaluation* project, which is affiliated with the National Institutes of Health, Project Number 8R01EB002124-03, and Principal Investigator, J. Michael Fitzpatrick, at Vanderbilt University, Nashville, TN.

2.1.3 Evaluating the Current Alignment

A set of special functions that are designed to evaluate the quality of the current alignment of the observations are called *similarity metrics* (or *objective functions*). During the registration procedure, they associate a numerical score to the current transformation estimate. Therefore, the goal of a registration algorithm can be interpreted as the optimization of such functions.

A subset of the similarity functions operates directly on the pixel or voxel intensities. These *intensity-based* methods calculate and compare various mathematical expressions (e.g.: joint density functions) using only the raw intensity values of the inputs. The number of observed samples (intensities) is determined by the size of the input data set, which can potentially be quite large. Over the past few decades, many such objective functions have been proposed [5, 42, 43, 55, 92]. Among these, a variety of methods are based on sound statistical principles including various maximum likelihood [37, 70], maximum mutual information [41, 84], minimum Kullback-Leibler divergence [10], minimum joint entropy [65] and maximum correlation ratio [54] methods. In this dissertation, we select a representative group of such statistical objective functions for further analysis. We explore their relative strengths and weaknesses, clarify the type of explicit and implicit assumptions they make, and examine their use of prior information. Through such an analysis, and some graphical representations of their solutions, we aim to facilitate a deeper and more intuitive understanding of their formulations.

Another group of objective functions relies on extracting key features (e.g., anatomical landmarks, fiducial markers, edge information, corners, etc.) from the input images and finding correspondences among them. *Feature- or landmark-based* approaches assume that a precise alignment of such a subset of the data samples is sufficient for aligning the rest of the input data. Using these methods, the raw intensity values need to be processed first, the key features need to be constructed or identified, and the matching criterion is then optimized to bring them into alignment. Contour- and point-based techniques [82, 66] are examples of this strategy, as well as registration methods that compare medialness properties of segmented anatomies [88]. Surface-based registration methods also belong to this category of alignment techniques [68, 18, 19, 75]. Instead of working directly with the volumetric data, these methods extract the surface of the imaged object and define a framework to register these lower-dimensional structures. One such approach, for instance, inflates the extracted cortical surface to the surface of the sphere and then calls for the alignment of these spheres instead of the highly folded cortical sheet [18, 19]. These techniques are popular in the analysis of functional images. For such applications, most of the information lies on the surface of the cortex and a careful alignment of the gyri and sulci is extremely important.

Once the key landmarks are identified, the optimization procedure usually proceeds much quicker in the case of feature- than the intensity-based methods. However,

major drawbacks of the former methods include the need to carefully plan the image acquisition protocols in advance, and the guaranteed presence and location of the trusted landmarks in the images. Such dependence on imaging and segmentation procedures can make it difficult to avoid the introduction of bias and (additional) errors into the subsequent registration algorithm. The feature-based solutions might also require some level of user interaction, which might not be desirable in certain medical applications. A retrospective performance analysis of some of the volumetric and surface-based methods is described in the work of West et al. [86].

The emphasis of the above summary of the registration problem is tailored somewhat towards our own analysis. We refer the reader for further details on the remaining varieties of currently used alignment approaches to the excellent review papers by Brown, Maintz, Roche and Zitová [5, 42, 43, 55, 92].

2.2 Subject-Atlas Registration

Another type of registration problem, which we can position directly between the uni-modal and multi-modal alignment problems, is the *subject-atlas* registration task. Although the objective functions and the transformations defined for this task may be exactly the same as described above, one of the data sets in the input might not correspond to a particular subject. Instead, it is a digital template that has been selected or constructed from a collection of previous observations, usually independently from the actual alignment task. In the following, we use the terms digital atlas, digital template and *reference* volume interchangeably.

In medical image analysis, subject-template registration procedures are commonly used when some type of previously obtained information is to be compared to or projected onto a newly examined data set. The prior knowledge is often accumulated through the analysis of numerous past observations. Depending on the information encoded in the atlas, a subject-atlas alignment might then facilitate a comparison between population and sample characteristics, or it might be used as a pre-processing step for segmentation studies.

We provide a more extensive definition of digital templates in Sec. 2.4, where we also demonstrate the tight link between registration and atlas construction.

2.3 Group-wise Alignment of Medical Images

In the previous sections, we mainly described registration scenarios that include one pair of image data sets. We can directly extend the notion of such *pair-wise* registration problems to *group-wise* problems. Accordingly, the task of aligning a collection of data sets is equivalent to establishing a homology among all the input images of

the set, where the number of input images is more than two. In this case, therefore, it is not one, but a group of transformations that needs to be identified in order to define complete correspondence among the group members. Given N volumes in the input data set, the exact number of transformations to be estimated has to be at least N, and at most $\mathbf{O}(N^2)$. The final answer depends on the algorithm specification: transformations can either be assigned to each member of the group, or alternatively, between each possible pair of the inputs. (For a special case, it is only $(N-1)$ transformations that are recovered. That occurs when one of the image set members is fixed and chosen as the template for the rest of the group.)

In the medical community, interest in aligning large sets of medical volumes has recently become prominent. This trend has been propelled by the emergence of new imaging modalities and an increased availability of fast computational resources and large storage units. Besides finding anatomical correspondences among each member of a set of observations, we can show how the results of these algorithms can also be used in order to construct digital atlases of the imaged anatomy. Models created from a set of observations can then be informative about the (ab)normal variations within or across various groups and could also be used as fixed references in future subject-to-atlas registration scenarios.

Besides many algorithmic details of the various group-wise registration methods, there is a particular property that helps to efficiently differentiate among them. This property defines how these registration approaches interpret and define the common coordinate frame where the proper correspondence of all the inputs is eventually defined.

Fixed Template-based Methods

For some applications, the desired common template is fixed and established in advance. Such a model can be, for example, defined as one particular member of the input data volumes [79, 24], or alternatively, the template can be identified as a special coordinate frame. The latter type of methods have been performed, for instance, with the construction of the Talairach anatomical coordinate system (defined by the identification of a set of key anatomical landmarks in all the data sets) [13, 24, 75, 69]. One of the major requirements of these methods is that all the input images need to be pre-processed for the matching landmarks to be reliably located. This is a time-consuming and potentially error-prone procedure that may skew the registration results. There is also a potential bias introduced into this registration framework, by claiming that a fixed template can sufficiently represent the whole group. It is possible to reduce the amount of such bias by updating the definition of the fixed model after all the images have been once aligned. For example, the mean of the aligned images can be computed and the group-wise registration repeated using the new mean as the template. Even though such model reassignment aims at reducing the bias introduced by the prior model, we cannot always ensure a non-biased implementation

of this process. In the case of anomalies present in the input, the registration results could be significantly distorted by an unfortunate choice of the initial model.

Online Template Definition

We also examine registration algorithms that do not use a pre-defined template, but instead generate one online, during the image alignment process [38, 45, 69, 79]. According to one such approach, class posterior joint statistics are aligned instead of intensity images. A template is defined such that it lies a minimum distance away from all of the samples and it also minimizes an entropy-based alignment score [31]. The process is iterated until the optimal alignment is found. Another approach follows a similar scheme, but it performs non-rigid alignment of the input scans using a minimum description length (MDL) criterion [45]. The reference frame is defined as the current mean or median of all the observed inputs and it is optimized such that, in an information theoretic sense, it best models all the group members.

Such formulations significantly reduce the amount of bias introduced into the registration framework. However, because of memory limitations and computational complexity, these algorithms can currently handle only a limited number of input volumes.

Template-free Approaches

In this section, we introduce the most general set of group-wise registration algorithms. In fact, these are the ones that mainly motivated our work. When using template-free registration approaches, there is no template or atlas defined either *a priori* or online. Instead, the input data sets are simultaneously aligned. Later, once alignment is achieved, the transformed input data set can be used to define the *central tendency* of the group.

Three different statistical approaches within this category define the image set registration problem by the generalization of a pair-wise alignment framework [63, 89]. Studholme et al. estimate the joint density function of all the inputs and construct a maximum likelihood-type similarity metric [63]. For computational ease, the input images are pre-segmented into a handful of anatomical classes. In [89], the authors introduce a registration framework based upon a higher dimensional generalization of a mutual information-type registration criterion. Yet another approach defines the image set registration problem by optimizing the average self-information (SI) metric [63], which is equivalent to minimizing over the sample joint entropy.

Many different variations of these objective functions have been previously proposed [3, 52, 64], but they all tend to have difficulties with larger numbers of input images. A drawback of the aforementioned approaches emerges in the form of computational complexity. By directly extending the pair-wise registration functions, their dimensionality linearly grows with respect to the number input images. Furthermore,

the number of samples required for a sufficiently accurate density estimate grows exponentially. Although existing probability density estimation methods might provide tight bounds on these quantities and allow for the simplification of these higher dimensional problems (e.g: [29]), the curse of dimensionality generally provides a great barrier for the usefulness and extensibility of these methods.

Last, but not least, we describe an algorithm that was first proposed in the machine learning and machine vision literature as a tool for handwritten digit recognition. The *congealing* data alignment technique calls for the simultaneous alignment of all the inputs [47, 48]. For an objective function it relies on the sum of voxel-wise entropies, which is then minimized. It only requires one dimensional density estimation even with an increasing number of inputs. Due to these attractive properties, we build on this congealing alignment framework when introducing our efficient population registration algorithm for 3D volumetric data sets. In addition to accommodating a potentially very large number of grayscale-valued medical data sets, it can be used in both uni-modal and truly multi-modal problems. A more detailed description of the method is presented in Chapter 4.

2.4 Building Digital Anatomical Atlases

As we mentioned in Section 2.2, in some scenarios, the registration process is defined with respect to a digital template and not another data set. In the following, we describe how these templates are selected or defined and how group-wise registration algorithms could be used in order to construct them.

While there is growing interest in the medical community towards using and creating digital anatomical atlases, the notion of such a template is only loosely defined in the literature [31, 38, 2, 79, 24, 50, 69]. As we briefly remarked earlier, many existing atlas-creation procedures either utilize the Talairach-based fixed coordinate system, or select one representative of the population as a target, align data sets to it, and then average over them. Others might use labeled data sets for segmentation purposes, or a set of other statistics gained from a set of registered data sets.

In general, all these approaches assume that the selected template, in one way or another, provides a fair representation of all other data sets belonging to the same group. While in practice there are numerous interpretations of digital atlases, they can all implicitly be closely related to a data compression or a modeling tool. In a sense, when creating a digital atlas, we want to encode information about the whole group with a reduced set of descriptors. Models of this sort can differ in their specificity and generalizability. At one extreme one data member acts as the model. That selection is very specific and does not generalize well to represent other observations; it does not encode information about the normal variability among the samples. At the other extreme all observations are kept. Although this model describes the entirety of observations perfectly, such a representation is very expensive in storage

requirements and it does not generalize well to new data sets.

We adopt a top-level definition of an atlas in the following way. An atlas is a representation that lies equal distances away from all the-observed data sets (in a particular high dimensional domain of the input images). Alternatively, it is a representation that *best* encodes the distinguishing properties of a group of samples according to a particular criterion. Defined this way, the atlas creation task can be very tightly associated with the group-wise registration problem. By aligning all the samples in the group, we hope to recover the *latent* common component present in all of them, to be able to also characterize more variable/non-standard features and to distinguish subtle differences between different groups.

Later, in Chapter 6, we propose the creation of a *mean brain* atlas via a simultaneous registration procedure, congealing. This process establishes the central tendency of a dataset by simultaneously warping all its components until a certain objective function is optimized. When this is achieved, distributional information is available about the underlying anatomy via the set of aligning transformations recovered by the algorithm. The congealing framework is also appealing as it does not pre-specify the number of underlying central tendencies. Thus, if our population contains more than one sub-population, it would be possible for the samples to converge towards more than one mode.

In addition to anatomical atlases, there is great interest in creating functional and combined functional-anatomical and also histological atlases[19].

2.4.1 Validation of Atlas Quality

Validating and verifying the accuracy of the resulting atlases is an interesting but also very complex task. Depending on the use of the atlas, multiple criteria can be established to assess its quality. Medical experts can determine the (anatomical) plausibility of the statistical volumes, and by comparing it to new images we can determine its generalizability (how well it can express non-observed images). Many times, it is necessary to identify specific applications and validate results within those. One of our applications, as it is described in Chapter 6, produces atlases of segmented labels by aligning MRI images of infant brains. These atlases are then used in an atlas-driven segmentation process. In this framework, registration quality can be indirectly evaluated by way of the resulting segmentations. Another framework for comparing the quality of such atlases can be found in [53].

2.5 Density and Entropy Estimation Strategies

Probability and density estimation both play a crucial role in our detailed analysis of the pair-wise registration methods in Chapter 3, as well as in defining an efficient

group-wise registration algorithm in Chapter 4. While we provide technical details on the type of estimators that we use in those chapters respectively, here we present a brief summary of non-parametric estimation methods that might serve as alternatives to the most commonly used techniques.

To facilitate the discussion to follow, we first define the term *entropy*. Intuitively, the word entropy refers to the amount of *disorder* or *uncertainty* of a system. In mathematical terms, however, the Shannon entropy measure of a discrete random variable X, with a probability mass function $p(x)$ is defined by [14]:

$$H(X) \equiv -\sum p(x) \log p(x) \ ,$$

or the average uncertainty in the random variable.

When estimating entropy, we can distinguish between *plug-in* and *non-plug-in* estimation methods. In the former case (and in the above definition), probability models are explicitly estimated and subsequently used in order to derive the entropy. In the latter, however, the entropy values are directly estimated without computing the corresponding probabilities. In certain scenarios – for example, when the available sample size is small in a higher dimensional problem – such a shortcut could provide a computationally efficient tool.

The most widely used plug-in estimators rely on the standard histogramming and Parzen Windowing approaches [17]. While histogramming offers speed and computational simplicity, Parzen Windowing provides a more principled framework for the computation of density derivatives, which is essential when dealing with gradient-based optimization tasks.

Non-plug-in estimators are not as widely used. Often, this is because they cannot be easily generalized to higher dimensional problems directly, or because partial derivatives with respect to a random variable are not trivial to compute. However, there does exist a certain set of the non-plug-in estimator methods that offers solutions for one or for both of these concerns.

For example, a competitive alternative to relative entropy (or Kullback Leibler-divergence) estimation has recently been introduced [33]. It extends the one-dimensional notion of the m-spacing estimators [76], which approximate the entropy by computing distances between order statistics. It is also possible to relate the sum of k-nearest-neighbor (kNN) distances to entropy measures [1]. The length of a minimum spanning tree (MST) is also directly related to the entropy of the data samples. This idea is incorporated into the entropic-graph estimators [27, 49]. Interestingly, for the MST-based approach, Sabuncu et al. have recently defined a framework for computing gradient-based directions, which can be used in efficient optimization algorithms [59].

Although we experimented with several of these methods, in our registration

framework we chose to use the EMMA-style[2] [77] iterative non-parametric approach for estimating entropies. This decision was chiefly influenced by the efficient gradient estimation characterizing this framework. We make use of both a smooth histogramming and a Parzen Windowing-style density estimator in its implementation. Further details of this framework are provided in Chapter 4.

2.6 Conclusion

In this chapter, we provided an overview of the medical image registration problem. We defined some key notions related to the registration task and introduced multiple scenarios where finding anatomical correspondences between/among the input data samples is crucial. We gave a brief review of the state of the art both in the pair-wise and group-wise registration domains. Furthermore, we pointed out the connection between the task of image alignment and the construction of digital anatomy templates, and briefly summarized the most frequently used entropy estimation techniques. In the upcoming chapter, we construct a unified information theoretic framework in order to explain the similarities and differences between a select group of statistical objective functions for registration.

[2]EMMA is a random but pronounceable subset of the letters in the words *Empirical entropy manipulation and analysis*[77]. This approach uses Parzen density estimators in a plug-in style.

Chapter 3

Statistical Methods for Multi-Modal, Pair-Wise Image Registration

In this chapter, we interpret a select group of multi-modal pair-wise registration methods in the context of a unified statistical and information theoretic framework. We clarify the implicit assumptions of each method, emphasizing their relative strengths and weaknesses, and provide diagrammatic interpretations to develop intuition. Additionally, we derive a new pair-wise registration framework motivated by our analysis. It combines the strength of currently used techniques and follows naturally from the analysis framework developed in the chapter.

3.1 Introduction

In the past two decades, numerous objective functions have been proposed to guide medical image registration. While the advantages of each function have been demonstrated individually, limited effort has been directed toward comparing all of them in a unified framework. In fact, we are only aware of a single overview of uni-modal registration methods – it covers a selective, strictly maximum likelihood-based summary accompanied by *functional* dependence between modalities [54].

In this chapter, we place a group of popular intensity-based pair-wise image registration techniques into a common statistical and information theoretic framework. By doing so, we are able to contrast them in terms of explicit and implicit assumptions, use of prior information, context-specific performance, and failure modes. In particular, we jointly analyze registration methods based on well-known statistical principles, such as maximum likelihood [37, 70], maximum mutual information [41, 84], minimum joint entropy [65], minimum Kullback-Leibler divergence [10] and maximum correlation ratio [54]. We also show why we prefer an information theoretic formulation of

the unifying framework rather than a maximum likelihood-based one. In addition to the theoretical analysis, we also provide a graphical representation for each method to facilitate a more intuitive understanding.

Following the results of the unified analysis, we formulate a novel pair-wise registration criterion that combines the strengths of previous approaches and we describe its performance on a set of particularly challenging alignment problems.

3.2 Notation and Key Concepts

The search for the optimal transformation that would establish proper correspondence between the input data sets is often cast within the classical parameter estimation framework. In this formulation, each unknown parameter is assumed to have a fixed value and is estimated using a maximum likelihood (ML) criterion. The popularity of the ML methods in parameter estimation can be explained by the fact that, in general, they have good convergence properties, they are relatively simple to implement, and as the sample size increases, they are asymptotically unbiased and efficient [72].

In the following, we first propose a general maximum likelihood framework and demonstrate how the classical ML problem formulation differs from the practical implementation of the approach. Then we show how likelihood functions can be viewed as estimates of, or approximations to, information theoretic measures. As this formalism facilitates the analysis of a wider range of registration techniques we use it to interpret the selected group of objective functions.

3.2.1 Notational Conventions

There exist a large number of notational conventions in the registration literature. In order to avoid confusion, we first specify the notation we use throughout the rest of this dissertation. While we select a multi-modal pair-wise alignment problem to demonstrate all the key concepts, it is relatively straightforward to generalize the analysis to group-wise registration problems.

Two *registered* input data sets $u(x)$ and $v(x)$ sampled on $x \in \Re^K$ represent two images of different modalities taken of the same underlying anatomy in a K-dimensional space. In practice, however, we do not observe the aligned images. Instead, they are offset by an unknown transformation. Thus we observe $u(x)$ and $v_o(x)$ (and not $v(x)$!) in which the latter is related to $v(x)$ by

$$v_o(x) = v(T^*(x)) \quad \text{or} \quad v(x) = v_o\left((T^*)^{-1}(x)\right), \tag{3.1}$$

where $T^* : \Re^K \rightarrow \Re^K$ is a bijective mapping corresponding to the unknown offsetting transformation[1]. The goal of registration is to find an estimate of an *aligning* trans-

[1]Technically speaking, $u(x)$ may have undergone some transformation as well, but without loss of

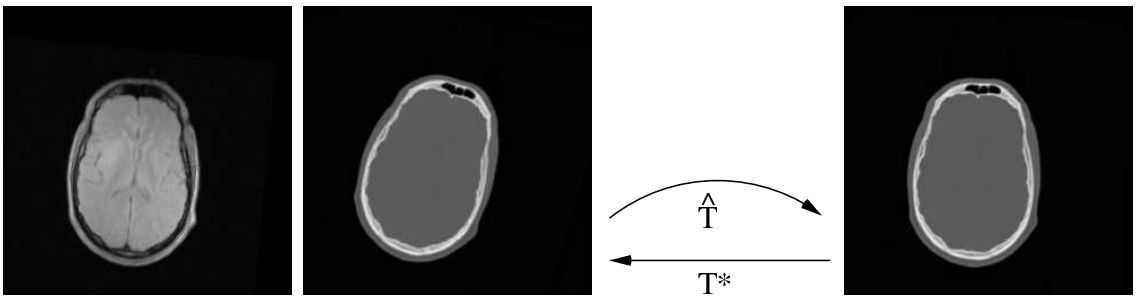


Figure 3-1: A 2D example of the registration problem. The *observed* input images are $u(x)$, an MR slice, and $v_o(x)$, a CT slice. $v(x)$ is the CT slice that is in correct alignment with the MRI slice. The unknown transformation that relates the observed CT data to the aligned image is T^*. The goal of the registration algorithm is to make $\hat{T}$ be the best estimate of $(T^*)^{-1}$.

formation $\hat{T}$ that best approximates the inverse of that transformation ($\hat{T} \approx (T^*)^{-1}$), usually by optimizing some objective function of the observed data sets.

Figure 3-1 demonstrates an example of an alignment scenario through a two-dimensional multi-modal pair-wise registration task. The input images are corresponding slices of an MRI and CT scan of the head that are initially mis-aligned, where we denote the MRI slice as u and the CT slice as v_o. In this scenario, we wish to apply a transformation to the CT slice to align it with the MRI. In other words, our task is to optimize $\hat{T}$ to best approximate the *inverse* of the underlying true aligning transformation T^*, so that $v_o(\hat{T}(x)) \approx v(x)$.

Intensity-based methods examine a set of intensity samples that are associated with corresponding spatial locations of the input data sets. Throughout our analysis (and consistent with general practice), spatial samples x_i are modeled as independent random draws from a uniformly distributed random variable X whose support is the coordinate system where the images are defined. Consequently, a property that is utilized by each of the pair-wise methods is that observed intensities $u(x_i)$ and $v_o(x_i)$ can be viewed as independent and identically distributed (*i.i.d.*) random variables, irrespective of spatial dependencies within the data. This follows from the fact that under fairly general conditions a *function* of an *i.i.d.* random variable is itself an *i.i.d.* random variable. Therefore, when we observe the input data sets, we can represent them as a sequence of joint measurements drawn *i.i.d.* from a particular joint distribution. In our work, we define the joint distribution that is assumed to produce the observed data sample pairs as the *source* distribution. We use parameter S in order to define its properties and write

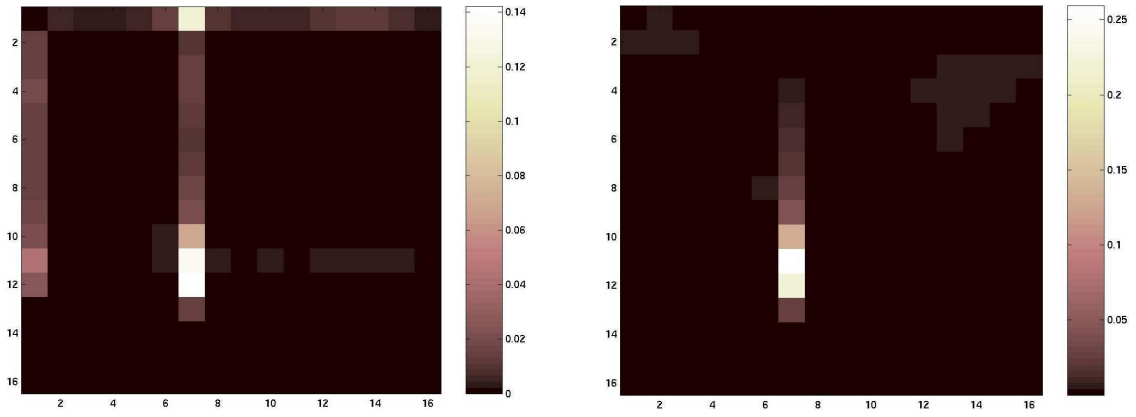
$$p_S = p\,(u, v; S)\,.$$

generality, we assume it has not. If there were some canonical coordinate frame (e.g. an anatomical atlas) by which to register the data sets one might consider transformations on $u(x)$ as well.

In most of our analysis, we also make the assumption that the distribution parameter is, in fact, a transformation. Thus the set of intensity sample pairs drawn from the source distribution p_S (or set of intensity sample pairs drawn from data sets that are related by transformation S) is indicated as:

$$\begin{aligned} \mathcal{Y}_S &\triangleq \{[u(x_1), v(S(x_1))], \ldots, [u(x_N), v(S(x_N))]\} \\ &= \{[u, v_S]_1, \ldots, [u, v_S]_N\}. \end{aligned} \tag{3.2}$$

It has been demonstrated that the joint statistics of these sample pairs are significantly different in the case of aligned and misaligned data sets [12]. Figure 3-2, for example, demonstrates two joint histograms constructed from intensity samples of the MR-CT pair displayed in Fig. 3-1. The plot of (a) was computed when the data sets were misaligned and (b) when they were registered. Comparing these histograms qualitatively, we see that in the misaligned configuration the histogram values are less structured and appear in more disconnected blobs than in the registered case.



(a) (b)

Figure 3-2: The joint histogram of the MR-CT image pair from Fig. 3-1 in (a) misaligned and (b) aligned configuration. Qualitatively, the joint statistics look more structured and less spread-out in the latter case.

Many statistical intensity-based algorithms aim to define a measure of such "structure" and optimize over it while others might define an ideal configuration for the joint statistics (by examining previous registration results) and intend to find the transformation that, by re-aligning the inputs, would best approximate it. In order to mathematically describe such registration formulations, we define another distribution. The *model* distribution is the joint distribution with which we estimate the source distribution from the observed intensity samples. Using M to parameterize this distribution, we write

$$p_M = p(u, v; M).$$

Interestingly, registration is an example of a statistical estimation problem in which *both* the source and the model probability distribution functions can be varied with

respect to an unknown transformation. In the analysis below, we show how the algorithms vary the source and/or the model distribution(s) in order to achieve the optimal alignment.

3.2.2 Links Between ML and Information Theory

In this section, we show two information theoretic interpretations of the ML criterion. The definition of this relationship allows us to analyze a larger set of statistical similarity metrics than would be possible if we only relied on the maximum likelihood formulation. First, however, we briefly remind the reader of the definition of two key information theoretic notions: the Shannon entropy and the Kullback-Leibler divergence [14].

The Shannon entropy measure of a discrete random variable X, with a probability mass function $p(x)$ is defined by:

$$H(X) \equiv -\sum_{x} p(x) \log p(x), \tag{3.3}$$

or the expected uncertainty in the random variable. In the current and upcoming chapters we use the notation $H(X)$ and $H(p)$ interchangeably. The Kullback-Leibler (KL) divergence, sometimes referred to as the relative entropy, is a measure of the difference between two probability distributions. Given two probability distributions p and q of the discrete random variable X, it is defined as

$$D(p||q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)} \quad . \tag{3.4}$$

This divergence measure is not a true metric as it is not a symmetric. It is a non-negative quantity that equals zero if and only if the two examined distributions are equivalent, if $p = q$.

Given the notational conventions and sampling assumption described in Sec. 3.2.1, the (normalized) log-likelihood function of the input image observations generated by source distribution p_S with respect to the modeling distribution p_M can be defined as:

$$\begin{aligned} \mathcal{L}_M(\mathcal{Y}_S) &= \frac{1}{N} \log\left(p\left([u, v_S]; M\right)\right) && (3.5) \\ &= \frac{1}{N} \log \prod_i \left(p\left([u, v_S]_i; M\right)\right) && (3.6) \\ &= \frac{1}{N} \sum_i \log\left(p\left([u, v_S]_i; M\right)\right), && (3.7) \end{aligned}$$

where N indicates the number of joint intensity samples that we observe from the examined input data. In Eq. 3.6 we use the *i.i.d.* assumption in order to express the

joint distribution as a product of individual sample distributions.

A relationship between such an ML formulation and information theory can be established. Although several explanations exist, below, we mention two of them. First, we may use the fact that the finite sample expectation of Eq. (3.7) approximates the sum of two information theoretic entities. Therefore, by taking the expectation with respect to the *source* distribution, we can write

$$E_{p_S}\left[\mathcal{L}_M\left(\mathcal{Y}_S\right)\right] = -\left[D\left(p_S||p_M\right) + H(p_S)\right], \tag{3.8}$$

where $H(p_S)$ is the entropy of the source distribution and $D\left(p_S||p_M\right)$ is the KL divergence between the source and model distribution functions. For a more detailed derivation of this relationship, see Appendix A.

Additionally, the same relationship can be established by using the Weak Law of Large Numbers. According to that, the limit of Eq.(3.7) in probability, as N grows large, is:

$$\mathcal{L}_M\left(\mathcal{Y}_S\right) \underset{N\to\infty}{\longrightarrow} -\left[D\left(p_S||p_M\right) + H(p_S)\right]. \tag{3.9}$$

Note that this limit can even be made stronger by assuming, for example, that the random variable $\log p\left(U,V\right)$ has finite variance; an assumption that holds in practice.

3.2.3 Differences Among Registration Approaches

According to our analysis, the critical differences between many registration methods can be explained by how they interpret and estimate the source and model distribution functions. More precisely, the distinguishing characteristics among the registration methods that we selected to analyze lie in

I. which distribution is viewed as the *source* (p_S) and which is viewed as the *model* (p_M),

II. how the model distribution is inferred from the observed data sets,

III. whether information-theoretic measures are utilized implicitly or explicitly, and

IV. how such measures are incorporated into the method.

Throughout this chapter, our goal is to explicitly point out these differences when discussing the selected group of alignment criteria. Understanding these issues gives some guidance as to which methods are appropriate for particular registration situations.

3.3 The ML Formulation of Registration

In this section we introduce the classical maximum likelihood (ML) formulation of the image registration problem. It is important to understand this framework, as it

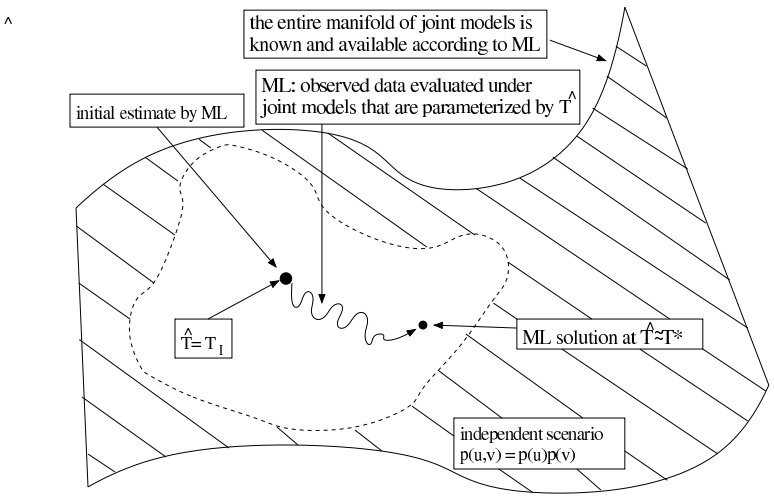


Figure 3-3: The space of joint distribution functions (the registration search space) parameterized by transformation $\hat{T}$. According to the classical ML approach, this entire space is known and available during the optimization procedure. The solution is defined at the transformation which maximizes the likelihood of the intensity pairs obtained from the input images. In the graphical display, the search starts at the identity transformation T_I and finishes at T^* where $\mathcal{L}_{\hat{T}}(\mathcal{Y}_{T^*})$ is maximized.

serves as the main building block of many currently available registration approaches. While we will see that practical issues generally preclude the use of a direct maximum likelihood implementation, analysis of the method is useful for comparison purposes. We point out, in the succeeding sections, how the currently used ML methods are different from the classical framework and some additional assumptions that are used to make them easy to implement.

Following the notation introduced in Sec.3.2.1 and referring to Eq.(3.1), we denote the observed input data sets as

$$\begin{aligned}\mathcal{Y}_{T^*} &= \{[u, v_{T^*}]_1, \ldots, [u, v_{T^*}]_N\} \\ &= \{[u(x_1), v(T^*(x_1))], \ldots, [u(x_N), v(T^*(x_N))]\} \\ &= \{[u(x_1), v_o(x_1)], \ldots, [u(x_N), v_o(x_N)]\}\end{aligned}$$

That is, the observed images $[u, v_o]$ are related by the unknown ground truth offsetting transformation T^*, which is considered to be the parameter describing the source distribution ($p_S = p_{T^*}$).

According to the ML criterion then, the optimal geometrical transformation that

explains these observations is

$$\begin{aligned} T_{\mathrm{ML}} &\equiv \arg\max_{\hat{T}} \mathcal{L}_{\hat{T}}\left(\mathcal{Y}_{T^*}\right) && (3.10)\\ &\approx \arg\min_{\hat{T}} \left[D\left(p_{T^*}||p_{\hat{T}}\right) + H\left(p_{T^*}\right)\right] && (3.11)\\ &= \arg\min_{\hat{T}} D\left(p_{T^*}||p_{\hat{T}}\right), && (3.12) \end{aligned}$$

where the model distribution is parameterized by the transformation $\hat{T}$ over which the optimization is computed. The approximation of Eq.(3.11) follows from the asymptotics demonstrated in Eq.(3.8) and Eq.(3.9). Also, in Eq.(3.12), the entropy term disappears as it is independent of the transformation parameter $\hat{T}$. Consequently, the classical ML approach selects the probability distribution $p(u, v; \hat{T})$ which is closest to the source distribution, $p(u, v; T^*)$, in the KL divergence sense.

In Figure 3-3, we depict the solution path associated with the classical maximum likelihood method. The graphic displays the space of joint probability distribution functions parameterized by transformation $\hat{T}$. According to the classical ML approach, this entire search space is known and available. We may start the optimization procedure at the identity transformation $\hat{T} = T_I$. That transformation is then modified during a local search mechanism in order to approach transformation T^*, which maximizes the ML criterion with respect to the currently observed images. Note that there exists a set of transformations in the optimization space that are indistinguishable under the registration criterion. We refer to them as *large* offsetting transformations, as applying them to the observations, the input data pair appears to be statistically independent. Joint distribution functions parameterized by such transformations are located outside of the dashed line on the display.

Finally, we point out a key difference between the classical ML framework and the rest of the pair-wise techniques (presented later). As described above, in the classical ML approach, while the *model* distribution is varied via the transformation $\hat{T}$, the observed input images – and thus the source distribution – remain fixed throughout the search process. Finding a globally optimal solution to Eq. (3.10) though requires that $p(u, v; \hat{T})$ be (pre)computed over all relative transformations $\hat{T}$ (see Figure 3-3). Although intuitive, this formalism, is impractical due to computational and memory limitations. While it might be feasible to use an optimization procedure that searches for a local optimum requiring the ability to produce $p(u, v; \hat{T})$ on demand, as far as we know, this alternative has not been tested or used. Instead, most of the registration methods, optimize an objective criterion by transforming the *observations* and thus modifying the source distribution. Consequently, the solution in their case needs to approximate the inverse of the underlying transformation, $(T^*)^{-1}$, and not T^* directly.

3.4 Unified Information Theoretic Analysis

In the subsequent analysis, six commonly used algorithms are shown to be approximations or estimates of the expected log-likelihood of observed intensities, the right hand side of Eq.(3.8). These are approximate maximum likelihood, KL divergence, joint entropy, iterated maximum likelihood, correlation ratio and mutual information. As foreshadowed in the previous section, these algorithms all make slightly different assumptions than the classical ML approach. Most importantly, instead of leaving the source distribution fixed, they apply the current transformation estimate to the observed input images prior to evaluating the objective criterion. The transformed observations, drawn from the varying source distribution, are denoted as:

$$\begin{aligned}
\mathcal{Y}_T &= \left\{ \left[u(x_1), v_\circ(\hat{T}(x_1)) \right], \ldots, \left[u(x_N), v_\circ(\hat{T}(x_N)) \right] \right\} \\
&= \left\{ \left[u(x_1), v(T^* \circ \hat{T}(x_1)) \right], \ldots, \left[u(x_N), v(T^* \circ \hat{T}(x_N)) \right] \right\} \\
&= \{ [u(x_1), v(T(x_1))], \ldots, [u(x_N), v(T(x_N))] \} \qquad (3.13) \\
&= \{ [u, v_T]_1, \ldots, [u, v_T]_N \}. \qquad (3.14)
\end{aligned}$$

In order to simplify the notation in our analysis, we introduce the following implicit definition. While, in practice, optimization is performed over $\hat{T}$ through $v_\circ(\hat{T}(x))$, we define the transformation parameter T to be the composition of two transformations: $T = (T^* \circ \hat{T})$. One is the unknown offsetting transformation T^* and the other is our actual current transformation estimate $\hat{T}$. In this manner T refers to the relative transformation applied to $v(x)$ rather than $v_\circ(x)$. While we express results on the implicit transformation, there are simple relationships which allow results to be expressed in terms of either T or $\hat{T}$ through the relation $v(T(x)) = v_\circ(T^* \circ \hat{T}(x))$. We emphasize that this convention is different from the one used in the classical ML analysis, where T is the transformation that directly parameterizes the model distribution function. For a summary of our slightly modified notational convention we refer the reader to Table 3.1.

T^*:	true offsetting transformation
$\hat{T}$:	transformation over which we optimize in practice; at correct alignment $\hat{T} \approx (T^*)^{-1}$
T:	$\left(T^* \circ \hat{T}\right)$, the transformation over which we optimize in our analysis; at correct alignment $T \approx T_I$ (T_I being the identity transformation)

Table 3.1: Notation of transformation parameters from Section 3.4 and beyond.

Accordingly, the normalized log-likelihood function of the observations generated by the changing source distribution and with respect to the model distribution p_M is

expressed as:

$$\mathcal{L}_M(\mathcal{Y}_T) = \frac{1}{N}\log\left(p\left([u, v_T]; M\right)\right) \tag{3.15}$$

$$= \frac{1}{N}\log\prod_i \left(p\left([u, v_T]_i; M\right)\right) \tag{3.16}$$

$$= \frac{1}{N}\sum_i \log\left(p\left([u, v_T]_i; M\right)\right). \tag{3.17}$$

The registration algorithms we analyze in this chapter differ in how they define either implicitly or explicitly the model distribution function p_M in order to improve the alignment criterion. The first subgroup of the objective functions relies on a fixed and known model distribution, the second computes the model joint distribution online, while the algorithms in the third group do not assume the existence of any target joint distribution function. Additionally, we also derive a new pair-wise registration framework that capitalizes on the strengths of the above mentioned algorithms. A brief organizational overview of how these objective functions are related to each other is presented in Fig. 3-4.

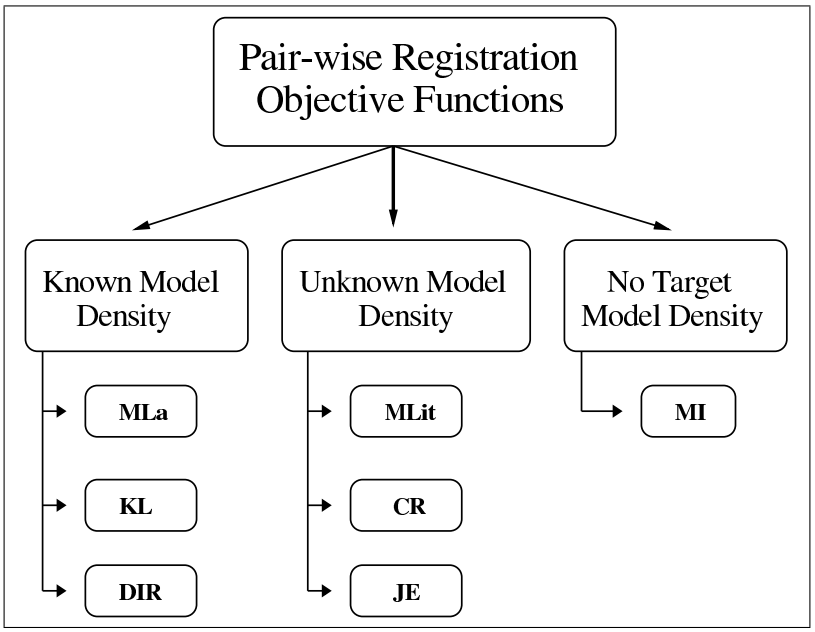


Figure 3-4: Organizational chart of the pair-wise registration objective functions that are discussed in Chapter 3. One subgroup relies on a fixed and known model distribution, another computes the model joint distribution online, while the algorithms in the third group do not assume the existence of a target joint distribution function to be modeled. The abbreviations of the indicated methods refer to: MLa – approximated maximum likelihood; KL – Kullback-Leibler divergence; DIR – ML approach with Dirichlet priors; MLit – iterated maximum likelihood; CR – correlation ratio; JE – joint entropy; MI – mutual information.

3.4.1 Using a Known Model Distribution

As we have already mentioned above, the availability of a model joint distribution function parameterized over *all* relative transformations T is generally infeasible. However, we might suppose that the model distribution parameters are fixed and known. In other words, we may assume that we have a model of the joint distribution of our data sets at one particular parameter setting, specifically when the modalities are registered. If we assume that this distribution is approximately the same for each image pair (of the same modality and taken of the same anatomy), we may estimate the model joint distribution, for example, from other *registered* data sets. Given a pair of aligned images u and v, where the offsetting transformation between them is the identity transformation T_I, we define the fixed model distribution function as $p^o(u,v) \equiv p\left(u,v;T_I\right)$. Any new set of corresponding intensity pairs can then be evaluated with respect to this.

The two objective functions discussed below, approximate maximum likelihood and Kullback-Leibler divergence, can both be related to the log-likelihood function written as:

$$\begin{aligned}
\mathcal{L}_{T_I}\left(\mathcal{Y}_T\right) &= \frac{1}{N}\log\left(p\left(\left[u,v_T\right];T_I\right)\right) \\
&= \frac{1}{N}\log\prod_i\left(p\left(\left[u,v_T\right]_i;T_I\right)\right) \\
&= \frac{1}{N}\sum_i\log\left(p\left(\left[u,v_T\right]_i;T_I\right)\right) \\
&= \frac{1}{N}\sum_i\log\left(p^o\left(\left[u,v_T\right]_i\right)\right) \qquad (3.18)
\end{aligned}$$

In order for a registration method to be successful using such a formulation, there are two underlying assumptions that need to hold:

I. it is feasible to estimate or learn a joint probability distribution model over the modalities of interest at the correct alignment[2], and

II. the resulting joint distribution model accurately captures the statistical properties of other unseen image pairs (presumably of the same anatomy and with the same modality pairing as the training set).

Approximate Maximum Likelihood (MLa)

This registration framework, which relies on a fixed and known model joint distribution function, was first suggested by Leventon and Grimson [37]. An identical formalism has been discussed more recently in [90]. In the following, we refer to the original formalism as the approximate maximum likelihood registration approach

[2]Assuming manual or other type of ground truth results are available from previous registration experiments.

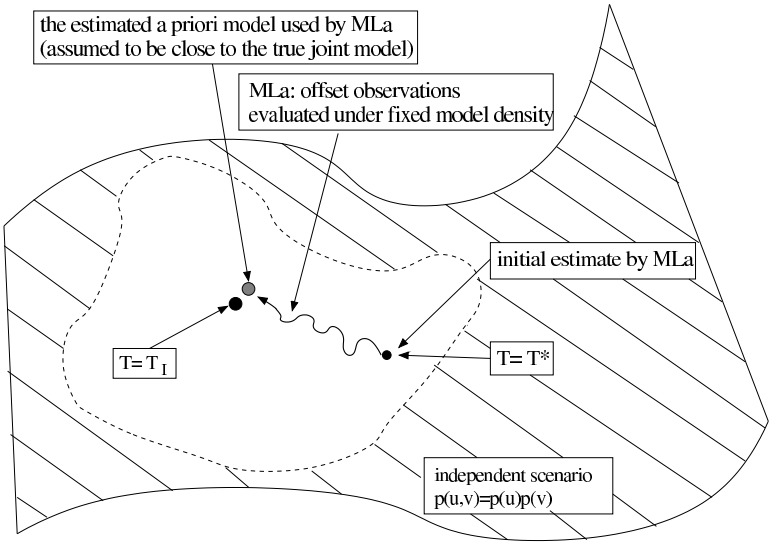


Figure 3-5: The approximate ML method, MLa, searches over the space of joint distributions parameterized by $T = \left(\hat{T} \circ T^*\right)$. It is at the identity transform T_I (or equivalently at $\hat{T} = T^*$) that the two input images are perfectly aligned. Starting with the observed input data samples (that are related via the unknown ground truth transformation T^*), the algorithm approaches the solution by evaluating the offset observations under a fixed model distribution.

(MLa).

Referring to Eq. (3.18), the MLa framework estimates the aligning transformation parameter T to be the transformation that maximizes an approximate likelihood criterion under the *a priori* distribution model p^o:

$$\begin{aligned} T_{\text{MLa}} &\equiv \arg\max_T \mathcal{L}_{T_I}\left(\mathcal{Y}_T\right) && (3.19)\\ &= \arg\max_T \sum \log p^o([u, v_T]_i) && (3.20)\\ &\approx \arg\min_T \left[D\left(p_T||p^o\right) + H\left(p_T\right)\right], && (3.21) \end{aligned}$$

where the approximation in Eq.(3.21) is the result of the asymptotics demonstrated in Eq.(3.8) and Eq.(3.9).

Contrary to Eq.(3.11), and as a consequence of manipulating the observations rather than the model distribution itself, the transformation T now influences both the entropy term and the KL divergence term on the right-hand side of Eq.(3.21).

We depict graphically a possible optimization path explored by the MLa algorithm in Figure 3-5. Starting with the observed input data samples (that are related via the unknown ground truth transformation T^*), the algorithm approaches the solution by evaluating the offset observations under a fixed model distribution. Note, in contrast

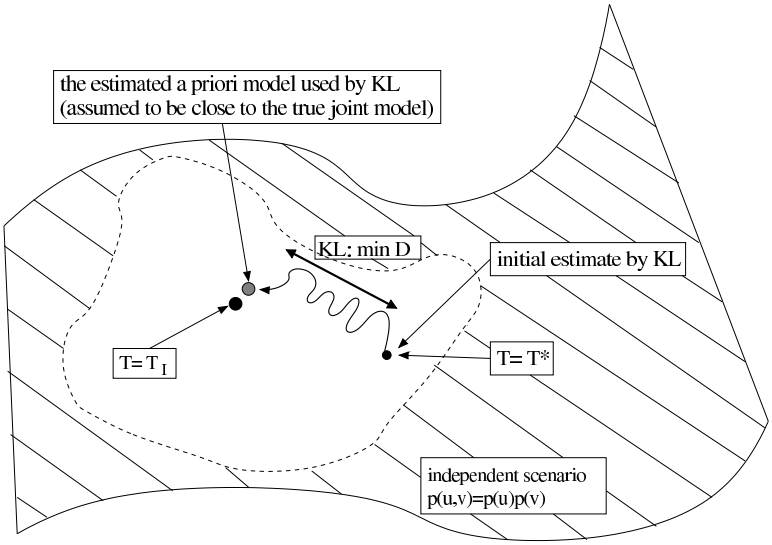


Figure 3-6: According to the KL registration framework, at each point of the search space a joint distribution function is estimated from the offset data pairs. The aligning transformation is located where the KL divergence (D) is minimized between that distribution estimate and a previously defined fixed model joint distribution.

to the classical ML method, the direction of the search path appears to be reversed. This is due to the modified assumptions explained at the beginning of Sec. 3.4.

Whereas the classical ML criterion ensured that the optimum occurs only when the *model* distribution under the hypothesized transformation agrees with the *source* distribution of the observations, it is possible that the terms in Eq.(3.21) conspire in such a way that transformations quite different from the aligning transformations yield strong local optima. Somewhat counter-intuitively, given a probability distribution function p, one may construct a distribution q such that typical draws from q have *higher likelihood* under p than typical draws from p, i.e. $-H(p) < -[H(q) + D(q\|p)]$. In that case, the optimization of the MLa criterion would not result in the desired high-quality alignment. Formally, this scenario can be explained by the information theoretic notion of *typicality* [14].

In the context of multi-modal registration, this behavior has also been observed empirically by Chung et al. [10] and it motivates their approach which is described below.

Minimizing Kullback-Leibler Divergence (KL)

Although not pursuing the analysis leading to Eq.(3.21), Chung *et al* suggested the use of KL divergence as a registration measure in order to align digital-subtraction angiography (DSA) and MR angiography (MRA) data sets [10]. Using the same modeling assumptions as in MLa (where we have a fixed distribution model p^o of the joint intensity data which is estimated from a set of previously registered data sets)

the authors optimize an objective function based on an estimate of a KL divergence term. The divergence is computed between the fixed model joint distribution and a probability model estimated from the transformed sets of observed data samples. Mathematically, this KL formulation of the registration problem is defined as

$$T_{\mathrm{KL}} \equiv \arg\min_{T} D\left(\hat{p}_T || p^o\right),$$

where $p^o = p(u, v; T_I)$ is constructed as in the MLa approach [37] from correctly registered data sets and $\hat{p}_T$ is a probability model estimated from the transformed sets of observed pixel intensities $\mathcal{Y}_T$ at a particular transformation T estimate. (This is in contrast to p_T from the MLa method in Eq. (3.21), where p_T indicates the true probability distribution of the input observations given parameter T.)

In Figure 3-6, we show a possible optimization path associated with the KL method. At each point of the search space (or at each estimate of the offsetting transformation T), a joint distribution is estimated from the transformed data pairs. The aligning transformation is located where the KL divergence (D) is minimized between the observed joint distribution estimate and the previously defined, fixed model distribution.

The previous methods approximate entropy and KL divergence terms via a likelihood function. Direct KL methods rely on numerical or Monte Carlo integration to evaluate KL divergence terms. Consequently, in addition to assumptions 1 and 2 from Sec. 3.4.1, this approach makes the following assumptions:

(KL-i) There is a reliable method for estimating $\hat{p}_T$ from transformed observations, and

(KL-ii) the KL divergence $D\left(p_T || p^o\right)$ can be accurately estimated via numerical or Monte Carlo integration of

$$\int\int \hat{p}\left(u, v; T\right) \log\left(\frac{\hat{p}\left(u, v; T\right)}{p^o\left(u, v\right)}\right) du dv \tag{3.22}$$

by substituting the estimate of $p\left(u, v; T\right)$, $\hat{p}\left(u, v; T\right)$, in the KL divergence integral.

This method has been demonstrated to be more robust with respect to (or less dependent on) the size of the sampling region and to have larger capture range than the ML (or the MI) approaches [10]. This robustness is partly explained by the informal discussion of typicality in the preceding section. That is, explicitly dropping the implicit entropy term in Eq.(3.21) at the cost of performing numerical integration avoids some local minima. Provided that both the distribution estimate and integration method are accurate, the KL divergence estimate is locally non-increasing as $\hat{T}$ approaches $(T^*)^{-1}$. This is supported by empirical comparisons in which KL did not exhibit some of the undesirable local extrema encountered in the MLa method [10].

As in the MLa method, the KL approach depends on a reliable joint probability model estimated from *aligned* data sets (i.e. assumption 1 in Sec. 3.4.1). That assumption could be quite restrictive. The available data sets for estimating such model densities typically have some misalignment which introduces modeling errors. While Chung et al. note that such models still yield good quality alignments when applied to new data sets, the model construction step could still introduce some amount of unwanted bias. For example, if the prior model is based upon only a few observations, outliers could skew the model distribution profile.

In relation to the ML and MLa methods, in the KL framework, both the samples ($[u, v_T]_i$) *and* the evaluation (source) distribution $\hat{p}(u, v; T)$ are being varied as a function of the transformation parameter T, while the algorithm approaches the static joint probability distribution model $p^o(u, v)$ constructed prior to the alignment procedure. Thus instead of evaluating the joint intensity pairs drawn from the transformed input data sets under the static model distribution, the KL approach re-estimates the source joint distribution function ($\hat{p}(u, v; T)$) at every iteration and uses that when evaluating the observations.

In summary, we analyzed two approaches to solving the registration problem. Both the MLa and the KL methods benefit from a fixed and known joint distribution model. Such information may increase the capture range of the algorithms, but we note that the accuracy of the solution is dependent on the accuracy of the prior joint distribution model. The optimizing transformation can only produce as good an alignment as that described by the model distribution. Any bias introduced into the model might also be reflected in the solution.

3.4.2 Unknown Model Distribution

In this section, we consider methods which discard the assumption of having a known and fixed model joint distribution of the image modalities to be registered. Instead, a framework is defined which necessitates an optimization over the transformation parameter T and a simultaneous estimation of the model joint distribution. As the model distribution is estimated online, it is also dependent on the estimates of the transformation T. In the following, we describe three different approaches to registration in this framework.

Joint Entropy

In the simplest scenario, similarly to the source distribution, we define the model distribution to be equal to p_T, and thus both $p_S = p_T$ and $p_M = p_T$. This criterion

modifies the one described in Eq.(3.17) in the following manner

$$\mathcal{L}_T(\mathcal{Y}_T) = \frac{1}{N}\sum_i \log\left(p\left([u, v_T]_i; T\right)\right). \tag{3.23}$$

Based upon the asymptotics demonstrated in Eq.(3.8) and Eq.(3.9), the optimization of this likelihood criterion leads to the minimization of an entropy measure:

$$\begin{aligned}
\arg\max_T \mathcal{L}_T(\mathcal{Y}_T) &= \arg\max_T \sum_i \log p([u, v_T]_i; T) && (3.24)\\
&\approx \arg\min_T \left[D\left(p_T||p_T\right) + H\left(p_T\right)\right] && (3.25)\\
&= \arg\min_T \left[H\left(p_T\right)\right]. && (3.26)
\end{aligned}$$

In words, the optimization of the likelihood criterion when both the source and the model densities directly depend on the transformation T is equivalent to the minimization of the entropy of the distribution p_T. As in practice the true source distribution needs to be estimated, the model distribution is indeed $p_M = \hat{p}_T$. Then we write:

$$\arg\max_T \mathcal{L}_T(\mathcal{Y}_T) \approx \arg\min_T \left[D\left(p_T||\hat{p}_T\right) + H\left(p_T\right)\right] \tag{3.27}$$

If we assume that the distribution estimate is sufficiently accurate and thus $\hat{p}_T \approx p_T$, then the KL-divergence term between them diminishes and we may formulate the objective criterion as:

$$T_{\text{JE}} \equiv \arg\min_T H\left(\hat{p}_T\right), \tag{3.28}$$

where $\hat{p}_T$, consistently with our previous notation, is a probability model estimated from the transformed sets of observed pixel intensities $\mathcal{Y}_T$ at a particular transformation T.

Again, both the observations and the distribution estimate used to evaluate the observations are dependent on T and change over the optimization procedure. As we demonstrate in Sec. 3.4.3, this registration objective is very similar to using mutual information. In fact, the emergence of joint entropy (JE) minimization preceded MI in the medical image registration literature [12]. However, as the latter proved to have a larger capture range and to be more robust, joint entropy alone is not widely applied.

Iterated Maximum Likelihood (MLit)

The intra-operative MR image alignment problem has recently been cast in an iterated maximum *a posteriori* (MAP) framework [70]. Timoner defined the objective function $\mathcal{F}$ in the form of optimizing $\mathcal{F}(u, v, T) = p(u, v|T)p(T)$ with respect to the transformation T. In his formulation, it is prior knowledge regarding the transformation parameters that is captured by the $p(T)$ term. For computational ease, the author also introduced an approximation to the joint distribution function under

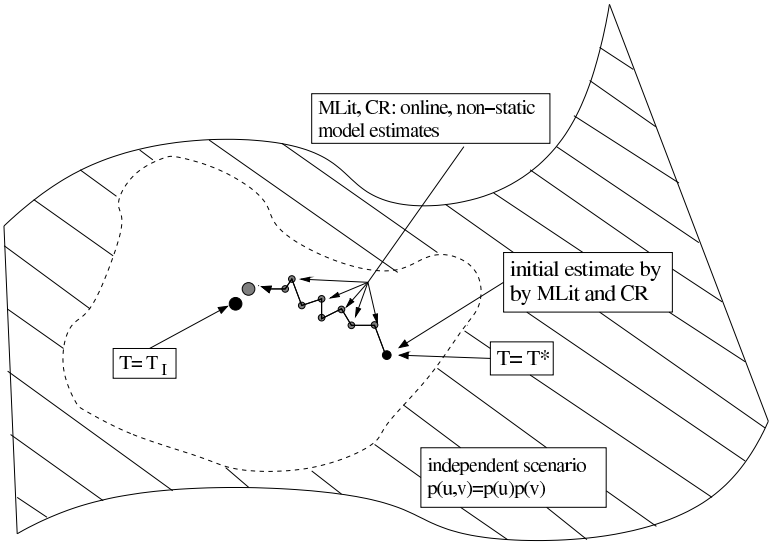


Figure 3-7: According to the MLit and the CR framework, at each point of the search space a joint distribution function is estimated using the most current transformation estimates. The transformation estimate T is updated in a way that the corresponding likelihood term is maximized.

which the observed joint intensity pairs are evaluated. Instead of the most up-to-date transformation estimate, he called for the utilization of an estimate from a previous iteration in the optimization procedure.

By neglecting the prior information on the transformation parameters, we can formulate this objective approach in an iterated fashion where each iteration of the algorithm consists of two basic steps. We may summarize these as:

(1) At iteration k, construct $\hat{p}\left(u, v; T^{(k)}\right)$ from samples $\mathcal{Y}_{T^{(k)}}$ that best estimates $p\left(u, v; T^{(k)}\right)$.

(2) Search locally over T to maximize $\mathcal{L}_{T^{(k)}}\left(\mathcal{Y}_T\right)$

Therefore, given the most recent estimate of the transformation $T^{(k)}$ as a model distribution, the next transformation estimate $T^{(k+1)}$, is defined to be the one that maximizes the normalized log likelihood over the varying source distribution p_T:

$$
\begin{aligned}
T^{(k+1)} &= \arg\max_T \mathcal{L}_{T^{(k)}}(\mathcal{Y}_T) && (3.29)\\
&= \arg\max_T \frac{1}{N} \log p\left([u, v_T]; T^{(k)}\right) && (3.30)\\
&= \arg\max_T \frac{1}{N} \sum_i \log p_{T^{(k)}}\left(u(x_i), v(T(x_i))\right) && (3.31)\\
&\approx \arg\min_T \left[D(p_T \| p_{T^{(k)}}) + H(p_T)\right]. && (3.32)
\end{aligned}
$$

We refer to this approach as the iterated ML objective function (MLit). As the

modelling distribution is explicitly estimated (resulting in $\hat{p}$), we express the optimal transformation at iteration $(k+1)$ as:

$$T_{\text{MLit}}^{(k+1)} \approx \arg\min_{T} \left[D\left(p_T || \hat{p}_{T^{(k)}}\right) + H\left(p_T\right) \right]. \tag{3.33}$$

Because of the re-use of model densities estimated in previous iterations, the MLit approach is straightforward to implement and promising registration results associated with this method are reported in [70]. In Figure 3-7, we show the optimization path that is covered by the MLit method. At each point of the search space a joint distribution is estimated using the most current set of parameter estimates. The transformation parameter estimates are updated in a way that the likelihood term is maximized.

In practice, the MLa and KL implementations also utilize an iterative optimization procedure. The current method differs from those in that the iteration is an explicit part of the optimization. This is due to the fact that the resulting transformation at each iteration is used to update the model distribution as well. Furthermore, during the maximization step transformations are searched for locally before re-estimating the joint distribution.

Correlation Ratio (CR)

Roche et al. [54] consider image registration in the case where there exists a functional relationship between the (multi-modal) input intensities when the data are aligned (i.e. $T = (T^*)^{-1}$). Specifically, for some function f and for all spatial coordinate index k

$$\begin{aligned} u(x_k) &= f(v(x_k)) + \epsilon_k \\ &= f(v_o((T^*)^{-1}(x_k))) + \epsilon_k \end{aligned} \tag{3.34}$$

where ϵ_k is additive stationary Gaussian noise. They derive conditions under which correlation ratio is an optimal similarity metric. The correlation ratio between u and v is defined as

$$\eta^2(u|v) = 1 - \frac{\text{Var}(u - \hat{f}(v))}{\text{Var}(u)}. \tag{3.35}$$

Thus, when using the correlation ratio (CR) method, it is assumed that:

(CR-i) the relationship between the intensities of the input modalities is described by Eq.(3.34), and that

(CR-ii) the noise ϵ_k is well modeled as additive stationary Gaussian noise.

This objective function has been well explained from a maximum likelihood perspective [55]. The joint probability distribution function of interest can be expressed in a product form $p(u, v; \gamma(T)) = p(v)p(u|v; \gamma(T))$, where it is a minor extension (aiding the subsequent analysis) to consider all T rather than only T^* as in [55]. The notation

$\gamma(T)$[3] makes explicit the notion that the probabilistic model comes from a parametric family (Gaussian in [54]) with parameters γ indexed by the transformation T. Note that it is possible to have $\gamma(T_1) = \gamma(T_2)$ for two very different transformations. A natural generalization of the CR method is to consider other parametric noise models (e.g. heavier tailed distributions or an outlier process). While the correlation ratio is no longer the optimal statistic, the basic idea for constructing the registration framework remains the same.

As $p(v)$ in the product does not depend on the transformation T, $p(u|v;\gamma)$ is optimized directly over γ. Instead of experimentally defining (and fixing) the joint probability model at the correct registration pose (as in the MLa and KL methods), the optimal probability distribution function is estimated online within the pre-selected parametric family. Finding the correct alignment of the input images is formulated as a coupled optimization task: a likelihood function over $p(u|v;\gamma(T))$ is alternately maximized with respect to T and γ. The necessary alternating optimization steps are equivalent to the optimization of Eq.(3.35), due to the following exponential relationship ([55]):

$$\begin{aligned} \eta^2(u|v(T)) &= 1 - \frac{1}{k} e^{2U(T)/N}, k = 2\pi \mathrm{Var}(u), \\ U(T) &= -\log \max_{\gamma} \; p(u|v;\gamma(T)). \end{aligned}$$

In order to relate this method to the other methods described herein, we replace the conditional distribution introduced above with the joint distribution $p(u, v; \gamma) = p_\gamma$.

Similarly to the iterative optimization search introduced in Sec. 3.4.2, we can then define the criteria of the CR iterations. The major difference here is that the model distribution estimation is formulated explicitly in a parametric manner, so the two steps of the iterations each correspond to the optimization of either the transformation or the distribution parameters. Following the objective function definition of MLit and denoting the distribution parameters, we can interpret the CR registration framework according to:

$$\Theta^{(k)} = \arg\max_{\Theta} \mathcal{L}_{T^{(k)},\Theta}\left(\mathcal{Y}_{T^{(k)}}\right), \tag{3.36}$$

and

$$\begin{aligned} T_{\mathrm{CR}}^{(k+1)} &= \arg\max_{T} \mathcal{L}_{T^{(k)},\Theta^{(k)}}\left(\mathcal{Y}_T\right) && (3.37) \\ &\approx \arg\min_{T} \left[D(p_T \| \hat{p}_{T^{(k)},\Theta^{(k)}}) + H(p_T)\right], && (3.38) \end{aligned}$$

[3]Note, in the original publication [55], the authors used the notation $\theta(T)$ to indicate the model distribution parameters indexed by T. In order to avoid notational conflicts with our conventions, we changed that to $\gamma(T)$.

where $T^{(k)}$ and $\Theta^{(k)}$ are both parameters obtained in the previous iteration and $\hat{p}_{T,\Theta}$ refers to the estimate of the joint distribution of the input image data sets, parameterized by transformation T and distribution parameters Θ. We emphasize that the Θ parameter here directly corresponds to the distribution parameters in an explicit parametric distribution estimation procedure. Note, no assumptions regarding parametric families for distribution estimation purposes were made in the previously described approaches.

From our analysis, it is clear that the objective function formulation using correlation ratio (Eq.(3.38)) is closely related to the MLit algorithm. Both approaches eliminate the use of prior joint distribution models. They approximate the model joint distribution online. A major difference between the two methods is that while the MLit technique applies a non-parametric distribution estimation method, the correlation ratio framework employs a particular parametric setting. Consequently, the latter approach implies two explicit optimization tasks. In fact, with the re-estimation requirement, it is possible to obtain a more accurate distribution model per case, but the sequential optimization of two individual functions could also get attracted to less favorable local solutions. It can be shown, that locally both of the methods converge to a joint entropy minimization.

We assign the same graphical display to the correlation ratio objective function as was introduced to the MLit procedure (see Fig. 3-7). As was emphasized before, the major difference between the two methods lies in the fact that the distribution optimization of the CR method is parametric, while in the case of MLit it is non-parametric. Otherwise the example optimization path through the registration search space of the two algorithms is the same.

3.4.3 No Target Model Distribution

There exist a set of registration methods that does not implicitly or explicitly aim to construct a target distribution. Instead, they formulate their objective criterions to favor scenarios that are located at maximum distance from the worst-case alignment. One example of such an approach is the popular mutual information alignment criterion. As we do not distinguish explicitly between a source and model distribution in this case, MI does not directly fit in the maximum likelihood framework. However, it can be connected to our unified information theoretic analysis through its KL divergence definition. It is exactly the existence of this group of registration formulations that prompted us to favor an information theoretic rather than an ML-based unified framework for our analysis. Otherwise this important group of registration approaches could not have been discussed.

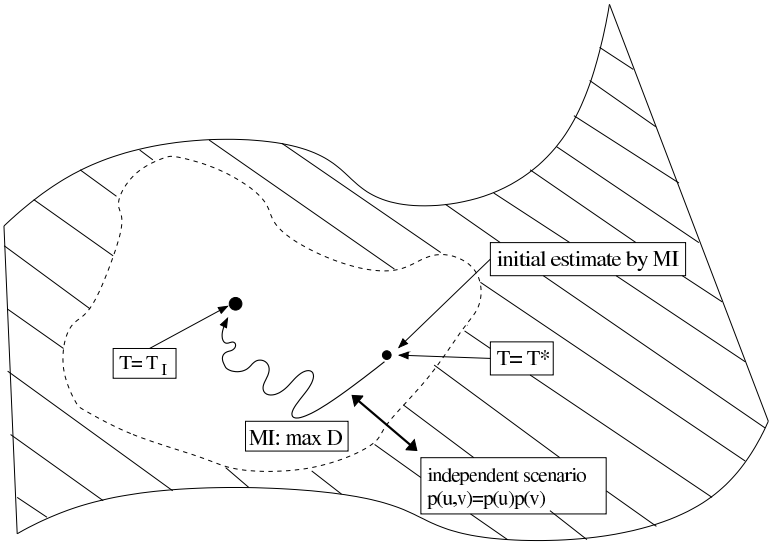


Figure 3-8: According to MI, the registration solution is located maximum KL divergence away from the worst-case, independent scenario, where the joint distribution is defined as the product of its marginals: $p(u, v; T) = p(u)p(v; T)$.

Mutual Information

As has been amply documented in the literature [14, 41, 51, 52, 84, 26], mutual information (MI) is a popular information theoretic objective criterion. It estimates the transformation parameter T by optimizing over the information theoretic quantity which, in the pair-wise image registration framework, one might define as:

$$\begin{aligned} I(u; v(T)) &= H(p(u)) + H(p(v(T))) - H(p(u, v(T))) \\ &= H(p(u)) + H(p_T(v)) - H(p_T(u, v)). \end{aligned} \tag{3.39}$$

Note that in the case of defining the mutual information metric, we need to examine both the marginal and the joint distribution functions of the input data sets. In order to distinguish between those, we will denote the marginals as $p(u)$ and $p_T(v)$ and the joint distribution as $p_T(u, v)$ (as opposed to simply p_T as otherwise introduced and used in the previous sections).

In addition to the entropy formulation, MI can also be expressed as a KL divergence measure [32]. In the registration scenario, we write

$$I(u; v(T)) = D(p_T(u, v) \| p(u)p_T(v)).$$

That is, mutual information is the KL divergence between the observed joint distribution term and the product of its marginals. Accordingly, the implicit assumption of MI methods is that

(MI-i) as $(T^* \circ \hat{T})$ diverges from T_I, or as we are getting farther away from the ideal registration pose, the joint intensities look less dependent.

This allows us to write the MI optimization problem as maximizing the distance from the *worst case* scenario, where the input images are completely independent:

$$T_{\text{MI}} \approx \arg\max_{T} D\left(\hat{p}_T(u,v) \| \hat{p}(u)\hat{p}_T(v)\right).$$

Recently, numerous variations on the mutual information metric have been introduced, such as, making it invariant to image overlap (e.g. normalized mutual information [65] and the related entropy correlation coefficient [41]), enhancing its robustness using additional image gradient information (gradient-augmented [51] and maximum distance-gradient magnitude-based mutual information [20]). In this dissertation, we do not list and analyze these, given that they operate with similar underlying principles.

As in the KL divergence alignment approach, both the samples and the evaluation densities are being simultaneously varied as a function of the transformation parameter T. However, instead of approaching a known joint distribution model according to KL divergence, the aim is to move the farthest away from the condition of statistical independence among the images, in the KL sense. This behavior is illustrated in Figure 3-8.

We mentioned in Sec. 3.4.2 that the minimization of joint entropy is closely related to the maximization of mutual information. Such a relationship is clearly visible if we consider the MI definition of Eq.(3.39). If T is restricted to the class of symplectic transformations (i.e. volume preserving), then $H(p(u))$ and $H(p_T(v))$ are invariant to T. In that case, maximization of MI is equivalent to minimization of the joint entropy term, $H(p_T(u,v))$, the presumption being that this quantity is minimized when $\hat{T} = (T^*)^{-1}$. In practice, however, it is extremely challenging to guarantee the symplectic behavior of the transformation functions and consequently the marginal entropy terms play a key role in the optimization.

Although MI has been one of the most popular objective functions in the multimodal registration literature, the existence of its global maxima about the point of correct registration has been only been observed and exploited empirically. To our knowledge, no sets of conditions have been previously established such that this global optimality criterion could be rigorously proved. In Appendix B, we provide a theoretical proof for the global optimality of the mutual information registration metric.

3.4.4 Summary

Considering the collection of pair-wise registration approaches discussed (a concise listing of them is provided in Table 3.2 and 3.3), we see that the MLa and KL divergence methods exploit prior information in the form of static joint distribution estimates over previously registered data. Subsequently, both make similar implicit assumptions regarding the behavior of joint intensity statistics as the transformation

estimate approaches the ideal alignment. In contrast, the joint entropy, the correlation ratio, the iterated maximum likelihood method and MI make no use of prior joint statistics. They instead, except for MI, estimate these during the search process. The mutual information-based approaches take a different approach. Instead of approaching a target joint distribution, they distance away from the most-undesirable scenario, when the input images are independent. In Table 3.2, we include all the pair-wise registration functions that we analyzed in the above section according to their unified information theoretic interpretation and we summarize in Table 3.3 some of their major assumptions that distinguish among them.

$$
\begin{aligned}
T_{\mathrm{ML}} &\approx \arg\min_{T} D\left(p_{T^*} || p_T\right) \\
T_{\mathrm{MLa}} &\approx \arg\min_{T} \left[D\left(p_T || p^{\circ}\right) + H\left(p_T\right)\right] \\
T_{\mathrm{KL}} &= \arg\min_{T} D\left(\hat{p}_T \| p^{\circ}\right) \\
T_{\mathrm{JE}} &= \arg\min_{T} H\left(\hat{p}_T\right) \\
T_{\mathrm{MLit}}^{(k+1)} &\approx \arg\min_{T} \left[D\left(p_T || \hat{p}_{T^{(k)}}\right) + H\left(p_T\right)\right] \\
T_{\mathrm{CR}}^{(k+1)} &\approx \arg\min_{T} \left[D(p_T || \hat{p}_{T^{(k)},\Theta^{(k)}}) + H(p_T)\right] \\
T_{\mathrm{MI}} &= \arg\max_{T} D\left(\hat{p}_T(u,v) \| \hat{p}(u)\hat{p}_T(v)\right)
\end{aligned}
$$

Table 3.2: The table summarizes the pair-wise registration formulas that are analyzed in this section positioned into the unified information theoretic framework.

	ML	**MLa**	**KL**	**JE**	**MLit**	**CR**	**MI**
source distribution (p_S)	p_{T^*}	p_T	p_T	p_T	p_T	p_T	p_T
model distribution (p_M)	p_T	p^o	p^o	p_T	$p_T^{(k)}$	$p_{T^{(k)},\Theta^{(k)}}$	-
model formation	fixed	fixed	fixed	on-line	on-line	on-line	-
term to optimize	D	H+D	D	H	H+D	H+D	D

Table 3.3: Concise comparison of the objective functions reformulated in the unified information theoretic framework.

3.5 A New Pair-wise Registration Method: Dirichlet Prior on Model Distribution

In this section, we formulate the pair-wise registration problem in a new way. We optimize over two parameters: the transformation parameter T and another set of variables Θ. The latter, similarly to its role in Sec.3.4.2, describes a set of distribution parameters. They contain additional information about the distribution profile, not characterized by the transformation parameters.

In contrast to the maximum likelihood view, where the parameters are assumed to be fixed but unknown, there exists another framework for the parameter estimation problem. According to the *Bayesian* interpretation, the parameters are viewed as random variables about which prior information (in the form of an *a priori* distribution model) is available. Observations in the form of the input images transform that prior into a posterior distribution, which may re-adjust our belief about the true value of the parameters to be estimated. Below, we define a prior distribution on the model distribution parameters Θ. Such information can be learned from previously registered image pairs.

3.5.1 Objective Function Definition

Given that we interpret Θ as a random variable, we can define a distribution over it, conveying prior knowledge about its values. If we do have such information, we might rewrite our registration objective function $\mathcal{F}$ in a *Maximum a Posteriori* (MAP) framework. This criterion is to be optimized over both the T and the Θ parameters:

$$\mathcal{F}(T, \Theta) = p([u, v_T]; T, \Theta) \tag{3.40}$$

Knowing the prior distribution over the parameter Θ, $p(\Theta)$, allows us to write:

$$\mathcal{F}(T, \Theta) = p([u, v_T]; T|\Theta)p(\Theta), \tag{3.41}$$

that is, we may express the joint distribution as the product of a conditional distribution $p([u, v_T]; T, \Theta)$ and the prior distribution $p(\Theta)$. (Note, it could also be natural to introduce a prior on the transformation parameters, if such information was available.)

We note that there is a very close connection between the MAP and ML formulations. The ML approach is, in fact, a special case of the MAP formulation of the problem. In the case of ML, we put a flat/uniform prior on the parameters. This might not be appropriate in all scenarios.

In the following we focus on a concrete scenario. We first assume that the joint distribution on image intensities is *i.i.d.* in space. As already mentioned, this assumption is made either implicitly or explicitly in most statistical approaches to intensity-based image registration. Further, we assume that the *i.i.d.* probability model is equivalent

to a multinomial distribution, which is an independent multi-trial model that uses the same probability mass function (*PMF*) for each trial. These PMF's are equivalent to normalized histograms which are widely used representations in statistical formulations of registration. In the JE and MI approaches, for example, the entropy of "images" is often estimated by histogramming the joint intensities, and calculating the entropy of the histogram. Aside from constant factors, this corresponds formally to using a plug-in entropy estimator that assumes a multinomial model, estimating the parameters of the model by histogramming, and calculating the entropy of the estimated model.

Returning to our development, the multinomial model with g basic outcomes may be parameterized by $\Theta = \{\theta_1, ..., \theta_g\}$ where $\theta_i \geq 0$ and $\sum_{i=1}^{g} \theta_i = 1$. These $theta_i$ correspond to normalized counts of the histograms described above. Let the random vector $Z = \{Z_1, ..., Z_g\}$ indicate how many times each event (joint occurrence of corresponding intensity values) occurs, then $\sum_{i=1}^{g} Z_i = N$, the number of independent trials. If we assume that the event probabilities are given by Θ, then the probability distribution of the random vector $Z \sim \text{Multinom}(\text{N}; \Theta)$ is given by

$$P(Z_1 = z_1, ..., Z_g = z_g) = \frac{N!}{\prod_{i=1}^{g} z_i!} \prod_{i=1}^{g} \theta_i^{z_i}. \tag{3.42}$$

According to this interpretation, Z corresponds to the event space of the joint intensity samples $[u, v_T]$ and N indicates the observed sample size. Such a representation is convenient, as prior information about the bin contents can be expressed by using Dirichlet distribution, the conjugate prior to a multinomial distribution.

Dirichlet distributions are multi-parameter generalizations of the Beta distribution. They define a distribution over distributions, thus the result of sampling a Dirichlet is a multinomial distribution in some discrete space. In the case where $\Theta = \{\theta_1, ..., \theta_g\}$ represents a probability distribution on the discrete space, the Dirichlet distribution over Θ is often written as

$$\text{Dirichlet}(\Theta; w) = \frac{1}{Z(w)} \prod_{i=1}^{g} \theta_i^{(w_i - 1)} \tag{3.43}$$

where $w = \{w_1, w_2, ..., w_g\}$ are the Dirichlet parameters and $\forall w_i > 0$. The normalization term is defined as

$$Z(w) = \frac{\prod_{i=1}^{g} \Gamma(w_i)}{\Gamma(\sum_{i=1}^{g} w_i)}, \tag{3.44}$$

where

$$\Gamma(w_i) = \int_0^\infty t^{w_i - 1} e^{-t} dt. \tag{3.45}$$

We, however, use another encoding of the distribution. We assign $w_i = \alpha m_i$, where

$\alpha > 0$ and $\sum_{i=1}^{g} m_i = 1$. Accordingly,

$$\text{Dirichlet}(\Theta; \alpha, M) = \frac{1}{Z(\alpha M)} \prod_{i=1}^{g} \theta_i^{(\alpha m_i - 1)} = \frac{\Gamma(\alpha)}{\prod_{i=1}^{g} \Gamma(\alpha m_i)} \prod_{i=1}^{g} \theta_i^{(\alpha m_i - 1)}. \tag{3.46}$$

This representation is more intuitive, as we can interpret $M = \{m_1, m_2, ..., m_g\}$ as a set of *base measures* which, it turns out, are also the mean value of Θ and α as a *precision* parameter showing how concentrated the distribution around M is. We can also think of α as the number of pseudo measurements observed to obtain M. The higher the former number is, the greater our confidence becomes in the values of M. When using a Dirichlet distribution, the expected value and the variance of the Θ parameters can be defined in closed form [21]. They are

$$\text{E}(\theta_i) = m_i \quad \text{and} \quad \text{Var}(\theta_i) = \frac{m_i(1 - m_i)}{\alpha(\alpha + 1)}. \tag{3.47}$$

Later we also need to compute the logarithm of this distribution which is equal to

$$\begin{aligned}
\log\left[\text{Dirichlet}(\Theta; \alpha, M)\right] &= \log\left[\frac{1}{Z(\alpha M)} \prod_{i=1}^{g} \theta_i^{(\alpha m_i - 1)}\right] && (3.48)\\
&= \log \prod_{i=1}^{g} \theta_i^{(\alpha m_i - 1)} - \log\left[Z(\alpha M)\right] && (3.49)\\
&= \sum_{i=1}^{g} \log(\theta_i^{(\alpha m_i - 1)}) - \log\left[Z(\alpha M)\right] && (3.50)\\
&= \sum_{i=1}^{g} (\alpha m_i - 1) \log \theta_i - \log\left[Z(\alpha M)\right]. && (3.51)
\end{aligned}$$

Thus, incorporating that assumption and normalizing the objective function in Eq.(3.41), we can optimize our proposed objective function as:

$$\begin{aligned}
\arg\max_{T,\Theta} \mathcal{F}(T, \Theta) &= \arg\max_{T,\Theta} \frac{1}{N} \log[p([u, v_T]; T|\Theta) p(\Theta)] && (3.52)\\
&= \arg\max_{T,\Theta} \frac{1}{N} [\log p([u, v_T]; T|\Theta) + \log p(\Theta)] && (3.53)\\
&= \arg\max_{T,\Theta} \frac{1}{N} \left[\log p([u, v_T]; T|\Theta) + \log \text{Dirichlet}(\Theta; \alpha M)\right] && (3.54)\\
&= \arg\max_{T,\Theta} \frac{1}{N} \left[\log p([u, v_T]; T|\Theta) + \sum_{i=1}^{g} (\alpha m_i - 1) \log \theta_i - \log Z(\alpha M)\right] && (3.55)
\end{aligned}$$

From an optimization point of view, we can discard the $\log Z(\alpha M)$ term, thus

$$\begin{aligned}
&\arg\max_{T,\Theta} \mathcal{F}(T,\Theta) = \\
&\quad = \arg\max_{T,\Theta} \frac{1}{N}\left[\log p([u,v_T];T|\Theta) + \sum_{i=1}^{g}(\alpha m_i - 1)\log\theta_i\right] && (3.56)\\
&\quad = \arg\max_{T,\Theta} \left[\frac{1}{N}\log p([u,v_T];T|\Theta) + \frac{1}{N}\sum_{i=1}^{g}(\alpha m_i - 1)\log\theta_i\right] && (3.57)\\
&\quad = \arg\max_{T,\Theta} \left[\frac{1}{N}\sum_{i=1}^{N}\log p(u(x_i),v(T(x_i));T|\Theta) + \frac{1}{N}\sum_{i=1}^{g}(\alpha m_i - 1)\log\theta_i\right] && (3.58)
\end{aligned}$$

As our goal is to solve a registration problem, we may choose to order the optimization of T and Θ. Instead of both the parameters, we require that only the optimal transformation T be returned. We define T_{DIR} to be the aligning transformation that optimizes the new objective criterion and write:

$$T_{\mathrm{DIR}} = \arg\max_{T}\left[\underbrace{\max_{\Theta}\left[\frac{1}{N}\sum_{i=1}^{N}\log p(u(x_i),v(T(x_i));T|\Theta) + \frac{1}{N}\sum_{i=1}^{g}(\alpha m_i - 1)\log\theta_i\right]}_{\hat{\Theta}_T}\right] \quad (3.59)$$

We define the distribution parameters that maximize the expression highlighted in Eq.(3.59) as $\hat{\Theta}_T$. This, in fact, corresponds to the MAP parameter estimate of the multinomial parameters given the image data and some value of T. Then

$$\begin{aligned}
T_{\mathrm{DIR}} &= \arg\max_{T}\left[\frac{1}{N}\sum_{i=1}^{N}\log p(u(x_i),v(T(x_i));T|\hat{\Theta}_T) + \frac{1}{N}\sum_{i=1}^{g}(\alpha m_i - 1)\log\hat{\theta}_{T_i}\right] && (3.60)\\
&\approx \arg\max_{T}\left[E_{p_T}\left[\log\hat{p}_{T,\hat{\Theta}_T}\right] + \frac{1}{N}\sum_{i=1}^{g}(\alpha m_i - 1)\log\hat{\theta}_{T_i}\right] && (3.61)\\
&= \arg\min_{T}\left[D(p_T||\hat{p}_{T,\hat{\Theta}_T}) + H(p_T) - \frac{1}{N}\sum_{i=1}^{g}(\alpha m_i - 1)\log\hat{\theta}_{T_i}\right], && (3.62)
\end{aligned}$$

where $\hat{p}_{T,\hat{\Theta}_T}$ is the estimated model joint distribution parameterized by T and $\hat{\Theta}_T$. The newly proposed objective function can be interpreted as the composition of a data- and a prior-related term. The former expresses discrepancies between the true source distribution and its estimated value, while the latter incorporates knowledge from previous correct alignments. As it might not be intuitive how that information influences the alignment criterion, in the following, we further manipulate the third term in Eq.(3.62).

The prior-related term in Eq.(3.62) can be expanded into a sum of two terms:

$$-\frac{1}{N}\sum_{i=1}^{g}(\alpha m_i - 1)\log\hat{\theta}_{T_i} = -\frac{\alpha}{N}\sum_{i=1}^{g} m_i \log\hat{\theta}_{T_i} + \frac{1}{N}\sum_{i=1}^{g}\log\hat{\theta}_{T_i}. \tag{3.63}$$

These two terms bear a close resemblance to the sample entropy and KL-divergence definitions. More precisely, if we assume that both the base parameters of the Dirichlet distribution $M = \{m_1, ..., m_g\}$ and the $\Theta = \{\theta_1, ..., \theta_g\}$ parameters represent normalized bin contents of a histogram encoding of probability mass functions $\mathcal{P}_M$ and $\mathcal{P}_{\hat{\Theta}_T}$, respectively, we could approximate the prior-related term through:

$$-\frac{\alpha}{N}\sum_{i=1}^{g} m_i \log\hat{\theta}_{T_i} + \frac{1}{N}\sum_{i=1}^{g}\log\hat{\theta}_{T_i} = \frac{\alpha}{N}\left[D(\mathcal{P}_M\|\mathcal{P}_{\hat{\Theta}_T}) + H(\mathcal{P}_M)\right] + \frac{1}{N}\sum_{i=1}^{g}\log\hat{\theta}_{T_i}.$$

Furthermore, if we denote a uniform probability distribution function by $\mathcal{P}_U$ where each of the g number of possible outcomes equals $\left(\frac{1}{g}\right)$, then

$$\begin{aligned}\frac{\alpha}{N}\left[D(\mathcal{P}_M\|\mathcal{P}_{\hat{\Theta}_T}) + H(\mathcal{P}_M)\right] + \frac{1}{N}\sum_{i=1}^{g}\log\hat{\theta}_{T_i} = \\ \frac{\alpha}{N}\left[D(\mathcal{P}_M\|\mathcal{P}_{\hat{\Theta}_T}) + H(\mathcal{P}_M)\right] - \frac{g}{N}\left[D(\mathcal{P}_U\|\mathcal{P}_{\hat{\Theta}_T}) + H(\mathcal{P}_U)\right].\end{aligned}$$

In summary then the objective function from Eq.(3.62) can be expressed as

$$\begin{aligned} T_{\text{DIR}} &\approx \arg\min_T \Big[D(p_T\|\hat{p}_{T,\hat{\Theta}_T}) + H(p_T) + \frac{\alpha}{N}\left[D(\mathcal{P}_M\|\mathcal{P}_{\hat{\Theta}_T}) + H(\mathcal{P}_M)\right] - \Big]. \\ &\qquad -\frac{g}{N}\left[D(\mathcal{P}_U\|\mathcal{P}_{\hat{\Theta}_T}) + H(\mathcal{P}_U)\right]\Big] \\ &= \arg\min_T \Big[D(p_T\|\hat{p}_{T,\hat{\Theta}_T}) + H(p_T) + \frac{\alpha}{N}D(\mathcal{P}_M\|\mathcal{P}_{\hat{\Theta}_T}) - \frac{g}{N}D(\mathcal{P}_U\|\mathcal{P}_{\hat{\Theta}_T})\Big]. \end{aligned} \tag{3.64}$$

Therefore, our new registration objective defined in Sec.3.5.1 can be interpreted as the weighted sum of four information theoretic terms. We refer to them as the *data* terms, the *prior* term and the estimation term. The first two terms, the data-related terms, indicate how well the true source distribution p_T is estimated by the model distribution given optimal distribution parameters $\hat{\Theta}_T$. The third term measures the KL-divergence between two probability mass functions over the parameters describing the *pseudo* and the current observations and the fourth term evaluates the KL-divergence between two other PMF's, the uniform and the one characterizing the parameters of the current observations. The last term can be interpreted as an alternative to a minimum joint entropy criterion. As the uniform distribution has the highest entropy among all, maximizing the KL-divergence from it is very similar to minimizing the entropy of the distribution.

As N is fixed and given by the number of the observed input intensity pairs, the weighting proportion depends solely on α, the *precision* parameter of the Dirichlet distribution. It is this value that determines how much weight is assigned to the *prior* term or in other words it ensures that the mode of the prior is centered on the previously observed statistics. That arrangement is intuitively reasonable. When α is high, the Dirichlet base counts are considered to originate from a large pool of previously observed, correctly aligned data sets; when α is low, prior observations of correct alignment are restricted to a smaller number of data sets. Thus in the case of a high α value, we have high confidence in the prior model, while in the case of a low α value, more emphasis is given to the characteristics of the currently analyzed data sets. It is interesting to note, that most often when one relies on fixed model densities, it is exactly this α value that is discarded. There is no notion about how many prior histograms or registered data sets have been observed in order to construct the known model distribution. We also point out that by discarding the prior information and assuming that the distribution estimation process is sufficiently accurate, the objective function is approximately equivalent to the joint entropy registration criterion.

Our new formulation of the registration problem is directly related to some recent registration efforts. As our theoretical analysis points out, it has also been established experimentally that using a fixed prior model distribution increases the capture range of the optimization search (inputs with larger offsets could be successfully aligned). The accuracy of the method, however, is biased by the quality of the model. Thus it is often desired to benefit from both a model-reliant and a model-free approach in order to guarantee both robustness and high alignment accuracy. Such ideas have been formulated by both Chung[4] and Guetter [25]. Chung et al. have proposed the sequential utilization of a KL-divergence and an MI term, while Guetter et al. incorporate the same two metrics into a simultaneous optimization framework. In both methods there is an arbitrary parameter that decides, respectively, when to switch between the two objective functions or how much weight to assign to each of them.

The above interpretation of our new metric following Eq.(3.64) presents a principled analysis about how one could balance the contributions of data and prior terms in an information theoretic framework in order to achieve more robust alignment solutions.

3.5.2 Preliminary probing experiments

In order to experimentally verify the previously claimed advantages of our novel pairwise registration algorithm, we designed a set of probing experiments. A probing experiment corresponds to the detailed characterization of an objective function with respect to certain transformation parameters. It helps to describe the capture range

[4]Private communications with Prof Albert Chung from The Hong Kong University of Science and Technology

(the interval over which the objective function does not contain any local optima besides the solution) and accuracy of the objective function. In our experiments we compared the behavior of four objective functions: joint entropy, negative mutual information, KL-divergence and our novel method. The first two of these methods only consider data information, the third one relies on previous registration results and our method incorporates both types of information.

The input data sets were 2D acquisitions of an MRI and an echo-planar MRI (EPI[5]) image (see Fig. 3-9). Historically, the registration of these two modalities has proved to be quite challenging because of the low resolution associated with the latter image [61].

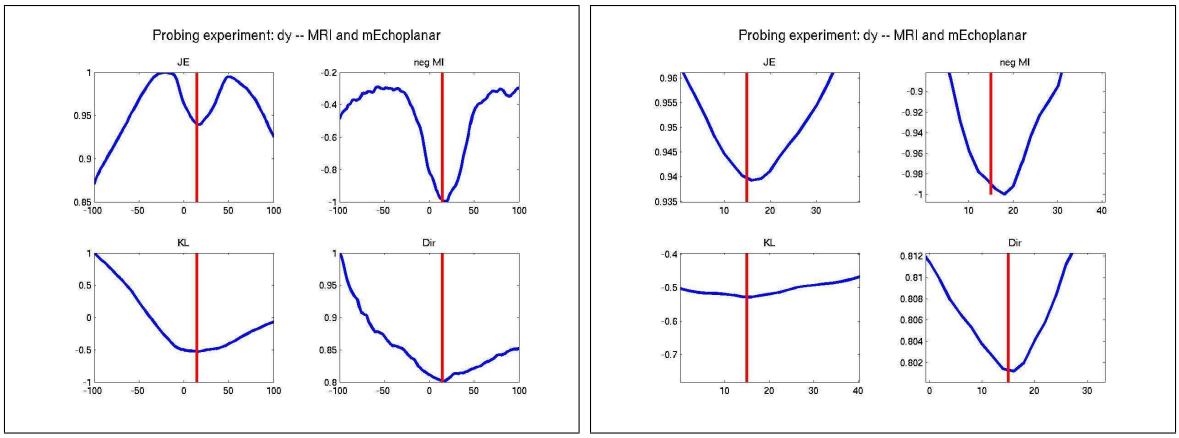
(a) (b)

Figure 3-9: 2D slices of a corresponding (a) MRI and (b) EPI data set pair. The probing experiments were run on these images.

We carried out the probing experiments in the y- (or vertical) direction. This is the parameter along which a strong local optimum occurs in the case of all the previously existing objective functions. In order to avoid any biases towards the zero solution, we offset the input EPI image by 15 mm along the probing direction. Thus the local optimum is expected to be located at this offset position – and not at zero – on the probing curves. The objective functions were all evaluated in the offset interval of [-100, 100] mm given 2mm step sizes. The results of the probing experiment are displayed in Fig. 3-10.

In the case of joint entropy, we find a close and precise local optimum corresponding to the offset solution location, but the capture range is not particularly wide. This means that beyond a narrow range of offset, JE demonstrates several local optima. In the case of negative MI, the capture range is just a bit wider. The KL objective function, as expected, increases the capture range. However, its accuracy in locating the offset optimal solution is not sufficient. In fact, around the expected local minimum the curve of the objective function is flat thus preventing the precise localization of the solution. The probing curve of our novel similarity metric demonstrates both large capture range and great accuracy. Thus relying on both previous registration results and the current observations this new metric is able to eliminate the undesired

[5]Echo planar imaging is a fast and efficient MRI technique which forms the basis of BOLD signal extraction in fMRI studies.



(a) Probing results (b) Close-up of the probing results

Figure 3-10: Probing results related to four different objective functions: joint entropy, MI, KL, our method (top-to-bottom, left-to-right).

local minimum solutions.

3.5.3 Connecting the Dirichlet Encoding to Other Prior Models

Finally, we diverge slightly from our main analysis. We draw similarities between the Dirichlet and other encodings of prior information on distribution parameters. Such an analysis facilitates a better understanding of the advantages of the Dirichlet encoding and it creates a tight link with other methods.

We start our analysis by showing that the maximum likelihood solution for the multinomial parameters Θ is equivalent to the histogrammed version of the observed intensity pairs drawn from the corresponding input images. Then, using these results, we demonstrate that the MAP estimate of the multinomial parameters (with a Dirichlet prior on them) is the histogrammed version of the pooled data, which is the combination of the currently observed samples and the hypothetical prior counts encoded by the Dirichlet distribution.

ML Solution for Multinomial Parameters

In this section, we rely on the same assumption that we made in Sec. **??**. Namely, the joint distribution of the observed samples is encoded with a histogram and we relate the normalized histogram bin contents to the parameters of a multinomial distribution over the random vector Z. Then the probability distribution of the random vector $Z \sim \text{Multinom}(\text{N}; \Theta)$ is given by

$$P(Z_1 = z_1, ..., Z_g = z_g) = \frac{N!}{\prod_{i=1}^{g} z_i!} \prod_{i=1}^{g} \theta_i^{z_i}. \tag{3.65}$$

Again, according to this interpretation, Z corresponds to the event space of the joint intensity samples $[u, v_T]$ and N indicates the observed sample size. If we want to then optimize the log version of this expression with respect to the Θ parameter, we write

$$\hat{\Theta} = \arg\max_{\Theta} \log \frac{N!}{\prod_{i=1}^{g} z_i!} \prod_{i=1}^{g} \theta_i^{z_i} \tag{3.66}$$

$$= \arg\max_{\Theta} \sum_{i=1}^{g} z_i \log \theta_i. \tag{3.67}$$

In other words, when searching for the maximum likelihood parameters of multinomial parameters, we need to compute the mode of the expression in Eq. (3.67) for all θ_i's. This formulation is very similar to that of the logarithm of the Dirichlet distribution which we formulated in Eq.(3.51). From probability theory we know that the mode of that expression is taken at $\left[\frac{\alpha m_i - 1}{\alpha - g}\right]$. Thus if we define $(\alpha m_i \equiv z_i + 1)$, the mode of Eq. (3.67) is found at

$$\hat{\theta}_i = \frac{\alpha m_i - 1}{\alpha - g} = \frac{(z_i + 1) - 1}{\sum_{i=1}^{g} z_i} = \frac{z_i}{\sum_{i=1}^{g} z_i}. \tag{3.68}$$

That is to say, the optimal θ_i parameter – in the maximum likelihood sense – is the one that can be computed by the number of corresponding counts normalized by the total number of counts. That is exactly the approximation that is utilized by the popular histogramming approach. Therefore, we can state that the maximum likelihood solution for the multinomial parameters is achieved by histogramming.

MAP Solution for Multinomial Parameters with a Dirichlet Prior

In this section we return to the MAP problem formulation that originated our analysis. Here, in order to find the optimal set of distribution parameters $\hat{\Theta}$ with a prior assigned to them, we have

$$\hat{\Theta} = \arg\max_{\Theta} \sum_{i=1}^{g} z_i \log \theta_i + \sum_{i=1}^{g} (\alpha m_i - 1) \log \theta_i \tag{3.69}$$

$$= \arg\max_{\Theta} \sum_{i=1}^{g} (z_i + \alpha m_i - 1) \log \theta_i \tag{3.70}$$

If we now define $\alpha' m'_i \equiv z_i + \alpha m_i$, then the above simplifies to

$$\hat{\theta}_i = \frac{\alpha' m'_i - 1}{\sum_{i=1}^{g} (\alpha' m'_i) - k} = \frac{z_i + \alpha_i - 1}{\sum_{i=1}^{g} (\alpha m_i) - k} = \frac{z_i + c_i}{\sum_{i=1}^{g} z_i + c_i}. \tag{3.71}$$

where $c_i = (\alpha m_i - 1)$ are *counting* parameters related to the *pseudo* counts of the Dirichlet distribution. That is to say, the optimal θ_i parameter – in the maximum a

posteriori sense – is the one that can be computed by the sum of the corresponding observed and pseudo counts normalized by the total number of observed and pseudo counts. In other words, in order to compute the optimal θ_i parameter, we need to pool together the actually observed and the pseudo counts and do histogramming on this merged collection of data samples.

Interestingly enough, this formulation forms a close relationship with another type of entropy-based registration algorithm. Sabuncu et al. introduced a registration technique based upon minimizing Renyi entropy, where the entropy measure is computed via a non-plug-in entropy estimator [59, 58]. This estimator is based upon constructing the EMST (Euclidean Minimum Spanning Tree) and using the edge length in that tree to approximate the entropy. According to their formulation, prior information is introduced into the framework by *pooling* together corresponding samples from the aligned (prior distribution model) and from the unaligned (to be registered) cases. Throughout the optimization, the model observations remain fixed and act as anchor points to bring the other samples into a more likely configuration. The reason why such an arrangement would provide a favorable solution has not been theoretically justified. Our formulation gives a proof for why such a method strives for the optimal solution.

Very recently, another account of relying on pooling of prior and current observations been published [71]. The authors use this technique to solve an MRI-CT multi-modal registration task.

3.6 Conclusion

We provided a unified statistical and information theoretic framework for comparing six well-known multi-modal image registration methods. We illustrated the underlying assumptions which distinguish them, and specifically, our investigation served to clarify the assumed behavior of joint intensity statistics as a function of transformation parameters. Additionally, we derived a novel pair-wise registration criterion that was motivated by our analysis. The flexibility of the new metric originates in the fact that although it is using a fixed joint distribution model to define its prior on the model joint distribution and to attract the initial registration estimates robustly towards the solution, the confidence level assigned to the prior distribution model might be adjusted. In the following chapter, we demonstrate how the above explained unified information theoretic analysis can be naturally extended to the group-wise registration scenario and we introduce our solution to this higher dimensional registration task.

Chapter 4

Group-wise Registration Methods & Congealing

In this chapter, we extend the notion of pair-wise image registration to that of group-wise alignment. The goal is to find correspondences among a whole group of data sets as opposed to just two of them. We demonstrate how the currently existing group-wise methods fit into the unified information theoretic framework discussed in Chapter 3 and we also establish a novel, unbiased and computationally-efficient framework for aligning populations of 3D medical images. Our framework is called stochastic congealing. *It brings into alignment a set of input volumes simultaneously, with every member of the population approaching the group's* central tendency *at the same time. In the second half of the chapter, we provide insights into implementation-related details of our algorithm.*

4.1 The Group-wise Registration Formulation

The introduction of new image modalities, and the increased availability of computational power and memory storage accessible to the medical community have facilitated the emergence of many interesting questions related to group-wise analysis. Besides examining individual images, the medical field has developed an interest in characterizing common and rare features within a population and comparing statistical characteristics of multiple sets of images. Most often, however, the set of data volumes is not directly comparable. The inputs might be taken over a longer period of time and possibly on multiple imaging scanners. For any type of group-wise study, we need to know how the corresponding anatomies are represented and where they are exactly located in the individual acquisitions. Therefore, group-wise registration of the input images frequently needs to precede any type of further analysis.

In Chapter 2, we gave a brief description of the currently available group-wise methods and hypothesized about how they could be used to construct digital anatom-

ical atlases. In this chapter, we show that our unified information theoretic framework can be naturally extended to describe a set of statistical group-wise registration methods. We focus on techniques that simultaneously align the input data sets as opposed to ones executing repeated pair-wise registration processes. We examine algorithms that treat the collection of input volumes as a whole and aim to fully benefit from the group's information. Namely, in the upcoming section we discuss the self-information metric [63], group-wise mutual information [52], the extensible information metric [89], and an MDL-based method [73]. We then introduce a novel efficient registration framework, *stochastic congealing*, that we developed in order to align large collections of both uni- and multi-modal medical data volumes.

4.1.1 Updated Notation and Definitions

When extending our analysis under the information theoretic framework to accommodate group-wise models, we introduce some notational changes that are necessitated by the higher number of input volumes.

We redefine the registration formulation to index multiple input images. Corresponding to our notation in Chapter 3, we denote the set of unaligned input images by $\mathcal{U} = \{u_1, u_2, ..., u_n\}$. We note that the u_i's indicate individual data sets that constitute the collection. We do not specify the nature of these observations. Unless otherwise noted, they could equally be from uni-modal or multi-modal, and from intra- or inter-subject image series. The goal of the group-wise registration task is to recover a set of n transformations, $\mathcal{T} = \{T_1, T_2, ..., T_n\}$, such that they best align the input volumes. Fig. 4-1 displays a schematic representation of the group-wise registration configuration. Note, that in the figure, we intentionally do not specify a common coordinate frame. In that way, the alignment can be done with respect to a pre-defined model, (only $(n-1)$ transformations would be recovered) or could be done without *a priori* specifying one.

As a basis for our group-wise analysis, we continue to refer to the maximum likelihood formulation introduced in Chapter 3. Here, however, we parameterize the joint distribution by a set of transformations as opposed to a single one. We indicate the set of observations sampled from source distribution $p_{\mathcal{S}}$ as:

$$\begin{aligned} \mathcal{Y}_S &\triangleq \{[u_1(S_1(x_1)), u_2(S_2(x_1)), \ldots, u_n(S_n(x_1))], \ldots && (4.1)\\ &\quad [u_1(S_1(x_N)), u_2(S_2(x_N)), \ldots, u_n(S_n(x_N))]\} && (4.2)\\ &= \{[u_{1S_1}, u_{2S_2}, \ldots, u_{nS_n}]_1, \ldots, [u_{1S_1}, u_{2S_2}, \ldots, u_{nS_n}]_N\}, && (4.3) \end{aligned}$$

where $\mathcal{S}$ is a set of transformations parameterizing the source distribution.

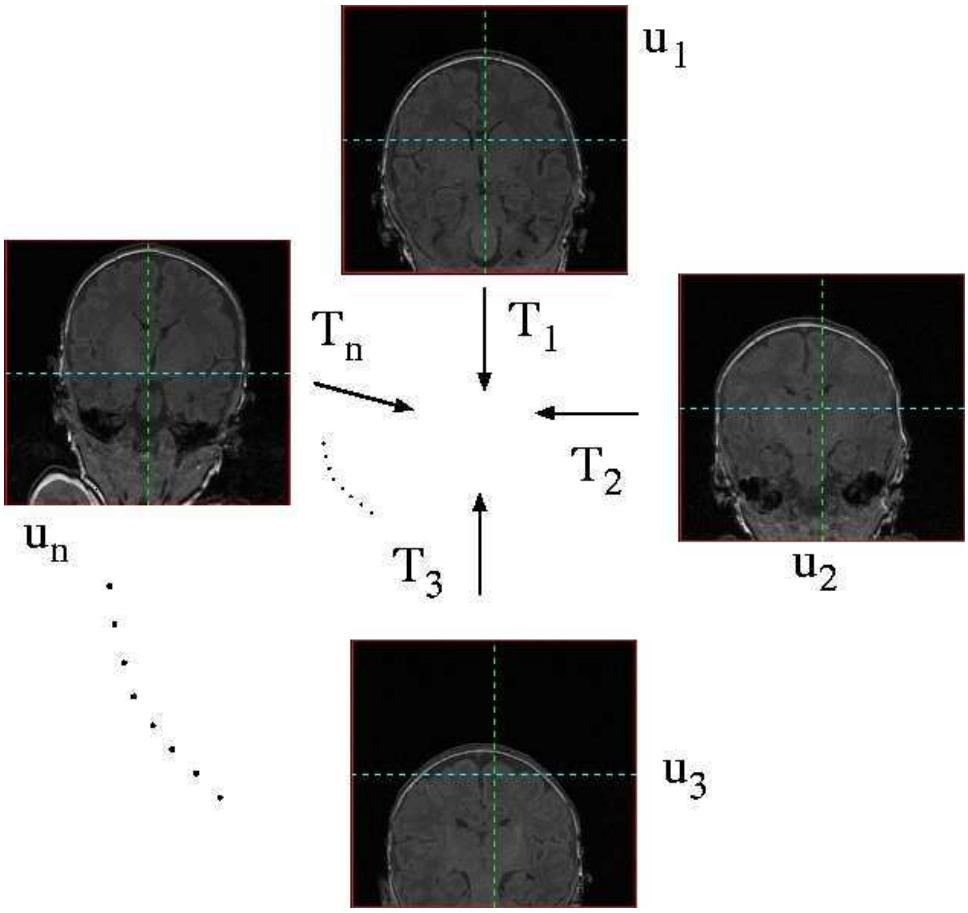


Figure 4-1: Example: group-wise registration configuration. Given n number of input images, n corresponding transformations need to be recovered in order to align the inputs. No specific model is defined for the common coordinate frame. In that way the alignment can be done with respect to a pre-defined model, or without specifying one.

Thus, the ML formulation of the group-wise registration problem becomes:

$$\begin{aligned} \mathcal{L}_{\mathcal{M}}(\mathcal{Y}_{\mathcal{S}}) &= \frac{1}{N}\log p([u_{1S_1}, u_{2S_2}, \ldots, u_{nS_n}]; \mathcal{M}) && (4.4)\\ &= \frac{1}{N}\log \prod_i p([u_{1S_1}, u_{2S_2}, \ldots, u_{nS_n}]_i; \mathcal{M}) && (4.5)\\ &= \frac{1}{N}\sum_i \log p([u_{1S_1}, u_{2S_2}, \ldots, u_{nS_n}]_i; \mathcal{M}), && (4.6) \end{aligned}$$

where N indicates the number of joint intensity tuples that we observe from the collection of input data sets, and $\mathcal{M}$ and $\mathcal{S}$ are two sets of transformations parameterizing the source and model densities respectively. In Eq.(4.5), we used the *i.i.d.* assumption in the spatial domain to write the joint probability distribution as a product of marginals.

Until now, the formulation agrees with the pair-wise registration case and we can still rely on the information theoretic link between ML and the entropy measure:

$$E_{p_{\mathcal{S}}}\left[\mathcal{L}_{\mathcal{M}}\left(\mathcal{Y}_{\mathcal{S}}\right)\right] \approx -\left[D\left(p_{\mathcal{S}}||p_{\mathcal{M}}\right) + H\left(p_{\mathcal{S}}\right)\right]. \tag{4.7}$$

4.1.2 Group-wise Self-Information

One recently introduced approach defines the population registration problem by optimizing the average self-information (SI) metric [63]:

$$\mathcal{T}_{\text{group-SI}} \equiv \arg\max_{\mathcal{T}} \frac{1}{N}\sum_{j=1}^{N} \log \hat{p}(u_1(T_1(x_j)), ..., u_n(T_n(x_j))),$$

where $\hat{p}$ indicates the estimate of the joint probability distribution of the transformed input data sets.

This objective function is a natural generalization of the minimization of pair-wise joint entropy introduced in Sec.3.4.2. Similarly to the corresponding analysis, we define the maximum likelihood problem with both the source and the model distribution directly parameterized by transformation parameters $\mathcal{T}$. Since in practice, our modeling distribution only estimates the true source distribution $p_S = p_T$, we express the modeling distribution as an approximation, $\hat{p}_T$. Thus

$$\arg\max_{T} \mathcal{L}_{\mathcal{T}}\left(\mathcal{Y}_{\mathcal{T}}\right) \approx \arg\min_{\mathcal{T}} \left[D\left(p_{\mathcal{T}}||\hat{p}_{\mathcal{T}}\right) + H\left(p_{\mathcal{T}}\right)\right]. \tag{4.8}$$

If we make the assumption that $\hat{p}_{\mathcal{T}}$ (the probability model estimated from the transformed set of observed intensities $\mathcal{Y}_{\mathcal{T}}$) provides sufficiently close approximation to the true source distribution, or if $\hat{p}_{\mathcal{T}} \approx p_{\mathcal{T}}$, Eq.(4.8) can be simplified to

$$\mathcal{T}_{\text{group-SI}} \approx \arg\min_{\mathcal{T}} \left[H\left(\hat{p}_{\mathcal{T}}\right)\right].$$

That establishes the link between the minimum joint entropy and the minimum self-information function.

4.1.3 Group-wise Mutual Information

There have also been several attempts to make use of the popular and widely-used mutual information similarity metric in higher dimensional applications [3, 52, 64]. Given more than two random variables, this metric can be formulated in many different ways. One of those definitions is the direct generalization of Eq.(3.39):

$$I(\mathcal{U}(\mathcal{T})) = I(u_1(T_1), u_2(T_2), ..., u_n(T_n)) \tag{4.9}$$

$$= \left[\sum_{i=1}^{n} H(u_i(T_i))\right] - H(u_1(T_1), u_2(T_2), ..., u_n(T_n)), \tag{4.10}$$

or in words, the sum of marginal entropies minus the overall joint entropy. This definition is easy to interpret in our unified framework. Using the KL interpretation, it can be shown:

$$\mathcal{T}_{\text{group-MI}} \approx \arg\max_{\mathcal{T}} D\left(\hat{p}_{\mathcal{T}}(u_1, ..., u_N) || \hat{p}_{T_1}(u_1) ... \hat{p}_{T_N}(u_N)\right),$$

where $\hat{p}_{\mathcal{T}}$ is an estimated probability distribution model.

Just as in the pair-wise case, mutual information aims to maximize the distance from the worst case scenario where the inputs are mutually independent (and their joint distribution equals to the product of their marginals). No explicit estimate of a target model joint distribution is constructed before or during the optimization.

A closely related information theoretic method introduces the *extensible information metric* as a group-wise similarity metric [89]. This function is equivalent to normalized mutual information and thus closely related to the above described mutual information.

The construction of the above two formulations is straightforward and intuitive. They align the input images as a group and they do not rely on or construct a specific target joint density throughout the registration iterations. Nevertheless, their implementation could become challenging. A linear increase in the number of the input arguments (data volumes) results in an exponential increase in the number of data samples required for distribution estimation. Even though there exist methods that provide upper bounds on entropy estimates using lower-dimensional computations (e.g.: [29]), and segmentation of the input images can be used to reduce the total size of the input data set, in the case of large data populations, the *curse of dimensionality* might severely restrict the performance of these approaches.

4.1.4 MDL-type registration

The work of Marsland and Twining interprets the population alignment problem as a data compression or model selection task. The purpose of statistical modeling is to discover regularities in the observed data. The success in finding such regularities can be measured by the codelength with which the data can be described. This is the rationale behind their Minimum Description Length (MDL) framework [45, 73, 74].

The authors introduce a simultaneous non-rigid registration framework. The objective function consists of a sum of entropy terms (description lengths) corresponding to the encoding of the data sets and the estimated aligning transformations. Besides the data fit term which describes the entropy of the reference volume and the sum of entropy of the discrepancy images, this method also optimizes over a complexity term which tends to prevent the usage of overly specific and excessively elaborate

models. As such a model complexity term is currently missing from our formulation, this MDL approach – though theoretically closely linked with a global maximum likelihood principle [23] – cannot be explained by our extended unified framework.

One drawback of the MDL-based approach is that the nature and the expected number of the central tendencies needs to be claimed and established in advance. In their most recent implementation, the authors selected a single volume computed as the mean or the median of all the observations. The current formulation is also restricted to using uni-modal data sets and comparing those by sum-of-square-differences. The mean intensity volume and such a quadratic objective function may not generalize well to multi-modal data populations. As we show in the upcoming section, our congealing-based method makes more general assumptions.

4.1.5 Congealing

In Chapter 2, we mentioned yet another template-free method called congealing which was originally introduced in the machine learning and machine vision literature [48]. Although its objective function, the sum of voxel-wise entropies, makes different independence assumptions from the rest of the algorithms, below we demonstrate how it still fits into our unified information theoretic analysis.

In Eq.(4.5), the group-wise version of the generalized ML framework, we used the *i.i.d.* sampling assumption to simplify the modelling. Just as in the pair-wise registration scenario, it means that the intensity samples drawn from the observed images are independently and identically distributed in space. For the sake of the congealing analysis, we modify that assumption. More specifically, we loosen it in the spatial domain and introduce an assumption in the imaging domain.

According to our new proposition, we assume independent but *not* identical distribution of the coordinate samples. That means that at each coordinate location x_i we need to estimate a different distribution p^i. If we again define the maximum likelihood problem with both the source and the model distribution directly parameterized by transformation parameters $\mathcal{T}$, the optimization task from Eq.(4.6) can be modified to:

$$\mathcal{L}_{\mathcal{T}}\left(\mathcal{Y}_{\mathcal{T}}\right) \approx \frac{1}{N}\sum_{i=1}^{N}\log p^i([u_{1T_1}, u_{2T_2}, \ldots, u_{nT_n}]_i; \mathcal{T}) \tag{4.11}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\log p^i(u_1(T_1(x_i)), \ldots, u_n(T_n(x_i))). \tag{4.12}$$

We emphasize that in Eq.(4.12) superscript i is associated both with the spatial samples and the local distribution estimates indicating that the latter are spatially varying.

Without any prior information, it is challenging to estimate the (p^i)'s merely from

the individual stack of observations. To facilitate the implementation, we can instead make another assumption: that the input images are independently and identically distributed in the input image domain. If that is the case (i.e. we replace the identicality assumption in the spatial coordinate space with an *i.i.d.* one in the input data domain), the joint densities of Eq.(4.12) can be written as a product, and the estimation task can be further simplified.

$$\mathcal{L}_{\mathcal{T}}\left(\mathcal{Y}_{\mathcal{T}}\right) \approx \frac{1}{N}\sum_{i=1}^{N}\log[p^i(u_1(T_1(x_i)))\ldots p^i(u_n(T_n(x_i)))] \tag{4.13}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n}\log p^i(u_j(T_j(x_i))). \tag{4.14}$$

The expression of the inner sum in Eq.(4.14) is closely related to sample entropy. That is, we can define an entropy estimator by $\hat{H}(z) = -\frac{1}{N_B}\sum_{z_i\in B}\ln\hat{P}(z_i)$, where Z is a random variable, B indicates data collections of Z of size N_B, z_i is an instance of Z and $\hat{P}$ is the estimated source distribution. Given that the true source distribution has to be estimated, the model distribution is set to $\hat{p}_{\mathcal{T}}$. The full objective function is equal to:

$$\begin{aligned}\arg\max_{\mathcal{T}}\mathcal{L}_{\mathcal{T}}\left(\mathcal{Y}_{\mathcal{T}}\right) &\approx \arg\min_{\mathcal{T}}\frac{n}{N}\sum_{i=1}^{N}H(\hat{p}^i_{\mathcal{T}}) \\ &= \arg\min_{\mathcal{T}}\sum_{i=1}^{N}H(\hat{p}^i_{\mathcal{T}}).\end{aligned}$$

That is, under the aforementioned assumptions (namely that the observed image intensities are independently but not identically distributed and that there is an *i.i.d.* relationship in the image domain) the group-wise ML criterion can be approximated by optimizing over the sum of one-dimensional voxel-wise entropies.

By observing the objective function, it is obvious that one of its great advantages is that even with a growing number of input image volumes to be registered it only requires the construction of one-dimensional distribution/entropy estimates, as opposed joint distributions with increasing dimensionality. Furthermore, the objective function does not single out a data set to represent a fixed template to which the rest of the inputs should be aligned. While recovering n different transformations corresponding to n input data sets, no target model or template is explicitly defined prior to or during the alignment.

As a brief summary, in this section we interpreted a set of group-wise objective functions in our extended unified information theoretic framework. The selected

methods all simultaneously align each member of the set towards an implicit reference frame. For an overview of these metrics, we provide Table 4.1.

$$\begin{aligned}
\mathcal{T}_{\text{group-SI}} &\approx \arg\min_{\mathcal{T}} H\left(\hat{p}_{\mathcal{T}}\right) \\
\mathcal{T}_{\text{group-MI}} &\approx \arg\max_{\mathcal{T}} D\left(\hat{p}_{\mathcal{T}}(u_1,...,u_N)||\hat{p}_{T_1}(u_1)...\hat{p}_{T_N}(u_N)\right) \\
\mathcal{T}_{\text{cong}} &= \arg\min_{\mathcal{T}} \sum_{i=1}^{N} H(\hat{p}_{\mathcal{T}}^{i})
\end{aligned}$$

Table 4.1: The table summarizes the group-wise registration formulas that are analyzed in this section positioned into the extended unified information theoretic framework.

4.2 Our Group-wise Registration Framework: *Stochastic Congealing*

We decided to build a new group-wise registration framework that could be applied to large collections of gray-scale valued medical image data sets. We chose to modify the *congealing*-style framework. Realizing its advantageous properties in earlier binary applications and its computational simplicity, we modified the original framework in a way that it enables the efficient registration of potentially very large sets of (multi-modal) grayscale-valued three-dimensional medical image data sets. We refer to our algorithm as *stochastic congealing* given the optimization method on which it relies. Our contribution lies in the implementation of a hierarchical stochastic gradient descent-based optimization method, which coupled with the one-dimensional entropy-based objective function, allows fast computation time even for data sets that contain more than a hundred input volumes. We also adapted our method to handle grayscale-valued images and designed a careful error analysis of the registration performance. In a set of experiments with both real and synthetic data sets, we quantitatively analyze the accuracy and the repeatability of our proposed algorithm. A top level summary of the algorithm is included in Algorithm 1 and in the rest of this chapter, we describe the key components of our registration algorithm. Then, in Chapter 5 and 6, we show the corresponding experimental results.

4.2.1 Favorable Properties

Given that there have been several group-wise registration methods proposed in the medical imaging community, we briefly summarize what the advantages of our stochastic congealing method are. Its favorable features include:

Algorithm 1 Top-level description of the congealing algorithm.

```
repeat
  for all Input data volumes do
    for all Randomly selected coordinate locations do
      Compute optimization update terms: partial derivatives of the current voxel-
      wise entropy with respect to the transformation components
    end for
    Sum the update terms
    Update the current transformation estimate
  end for
  Normalize the current set of transformations
until Convergence
```

- Computational simplicity: only the construction of one-dimensional distribution functions is required instead of higher-dimensional ones.

- One or more statistical model(s) of the central tendency of a set of brain volumes is derived from the data, rather than chosen *a priori* by the atlas creator.

- Local minima in the registration procedures, which can plague methods that align one brain volume at a time to a preselected standard, are often avoided by congealing. In effect, the ensemble of brain volumes provide a smoother optimization landscape for warping than the single reference scan provides in other registration methods [47]. One simple example (see Fig. 4-2) to illustrate that claim would be the alignment of hand images. In the case of pair-wise registration, when the initial offset of the inputs is large, the algorithm can easily get trapped in local optima situations where the non-corresponding figures on the input images are registered. With a larger pool of input images, the group-wise central tendency would dominate, thus individual outliers have smaller input on the final outcome.

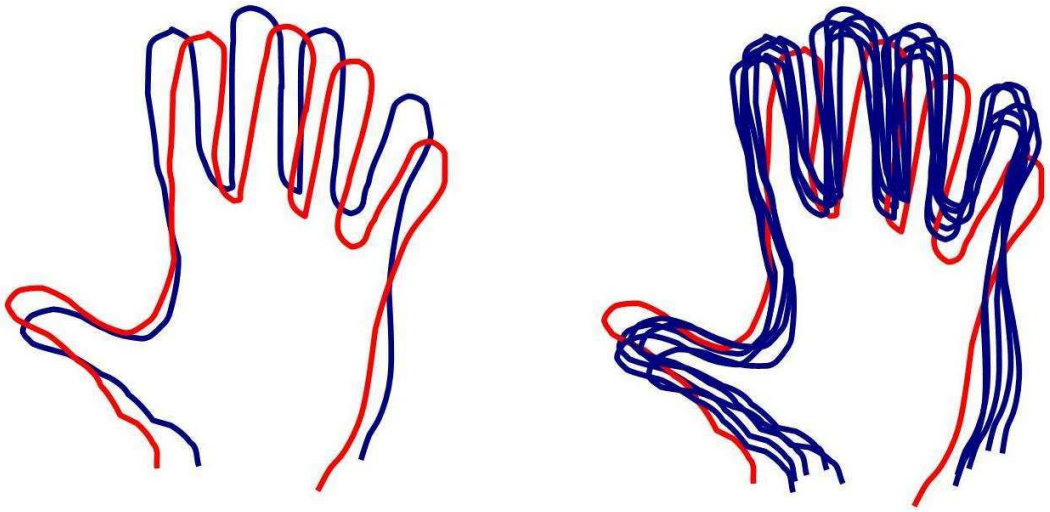
- No prior specification of the number of central tendencies is necessary. It is thus possible to identify multiple central tendencies in the congealed data set.

- No pre-processing, hand-labeling or pre-alignment of the data sets is required.

- The framework also allows for an easy extension concerning how to align a newly observed data set to a previously congealed data set without having to re-rerun the group-wise alignment process.

- Besides the uni-modal case, multi-modal data sets can also be aligned (given some modifications specified in the voxel-wise entropy estimation step).

These properties and their significance in specific registration applications are explained in more details and are experimentally demonstrated in Chapters 5 and 6.



(a) (b)

Figure 4-2: Unaligned set of hand outlines. Registration algorithms in the (a) pairwise scenario can easily get trapped in a local minimum situation while in the (b) group-wise scenario outliers can be more robustly accommodated for.

4.2.2 The Objective Function

The objective function of the stochastic congealing framework is the sum of voxel-wise entropies. In other words, a particular alignment is evaluated by computing the entropy of image intensities drawn from the same spatial coordinate location from all the input data. This is in contrast with the popular mutual information or entropy methods where joint entropy is computed *within* the input images [41, 63, 84].

In mathematical terms, if we denote the collection of n input images as $\mathcal{U} := \{u_1, u_2, ..., u_n\}$ and a set of n corresponding transformations $\mathcal{T} := \{T_1, T_2, ..., T_n\}$, the corresponding objective function $\mathcal{F}$ to be minimized is:

$$f(\mathcal{U}, \mathcal{T}) = f(T_1(u_1), ..., T_n(u_n)) = \sum_{i=1}^{N} H(\hat{p}(\mathcal{U}(\mathcal{T}(\mathbf{x_i})))),$$

where $\mathbf{x_i} \in \mathcal{R}^3$ indicates a particular coordinate location in the spatial coordinate system, H is the Shannon entropy and N to the total number of voxel locations in the image coordinate system.

4.2.3 Handling Grayscale Intensities

Many group-wise registration approaches, including previous implementations of congealing-based alignment, operate on binary values or segmented versions of the input data sets. The reason why they require such pre-processing of the inputs is that computationally it is more efficient to operate on these reduced-sized data volumes. Moreover,

with the segmentation process, it is often possible to eliminate noise and imaging artifacts that are present in the initially observed raw data.

In our framework, we use full-range grayscale intensity values. We use datasets that are either in the 8-bit or in the 16-bit data range. This provides a smoother search space for our gradient-based optimization as the intensity values do not change as dramatically as the binary or segmentation labels would. Even though pre-processing noise might be present in the data, our algorithm is robust enough to handle bias field corruption or other sources of imaging artifacts without having to have the input images segmented. Given that the input images are treated as a group, the noise factors tend to cancel out, and they do not have a significant impact on our registration results. Indeed, we found that working with a large scale of intensity values provides better entropy and distribution estimates. 5, we describe successful registration experiments even on challenging data sets, and we also compare our alignment quality with the results of an algorithm using binary inputs.

4.2.4 The Multi-Resolution Framework

It is widely known in the registration literature that optimization algorithms can easily become trapped in local minima [5, 42, 43, 55, 92]. To avoid such a scenario, one often constructs a multi-resolution framework where the processing of the data sets starts at a down-sampled and smoothed (low resolution) level and is refined during the higher resolution iterations. Not only can this framework improve the optimization performance, it also increases computation speed and memory usage efficiency. The number of hierarchy levels is mostly dependent on the quality and the original size of the input images. The higher these indicators, the higher the number of the processing levels. We also implemented such a multi-resolution framework for our group-wise registration method. In the case of the experiments presented in Chapter 5, it was sufficient to use a maximum of three levels.

4.2.5 Affine Transformations

In the lower levels of the processing pyramid, we use twelve-parameter affine transformations in order to align our input volumes. Given that T_k ($k = \{1, ..., n\}$) is a 3D affine transformation, it encodes rotation, scaling, shearing and displacement components. We might decompose the transformation into a displacement and an affine transformation component: $T_i = \{D_i, K_i\}$. The latter consists of rotation, anisotropic scaling and shearing. Our convention orders the transformation components as rotation, scaling and shearing followed by the displacement term. Note, while this ordering is arbitrary, it is important to follow it consistently, as the transformation components, in general, do not commute.

4.2.6 Affine Normalization

When simultaneously updating the transformations corresponding to each member of the input data set, a transformation component, common to all the estimated ones, may be sustained. The following simple example demonstrates this phenomenon. Let's compose all of our final transformation estimates by a random offsetting transformation. Even with this set of perturbed transformations, the same quality of alignment would be obtained and the same value of the objective function would be returned as before the perturbation. In an extreme situation, when this *perturbation* is too large, the input data sets could collectively drift off of the display producing an undesirable registration solution. To prevent such a scenario, we define a normalization step that is executed at the end of each iteration of the algorithm. We require that the average transformation at all spatial coordinate locations be the identity, T_{I}. If x_i corresponds to a particular coordinate location, T_o represents the unknown transformation that would guarantee such a criterion.

$$x_i = T_o \left[\frac{1}{N} \left(\sum_{j=1}^{N} T_j x_i \right) \right]. \tag{4.15}$$

Because the sum can be re-arranged, the affine regularizing transformation T_o is exactly the inverse of the average transformations:

$$T_{\mathrm{I}} = T_o \left[\frac{1}{N} \sum_{j=1}^{N} T_j \right] \quad \rightarrow \quad T_o = \left[\frac{1}{N} \sum_{j=1}^{N} T_j \right]^{-1}. \tag{4.16}$$

This update guarantees that the average movement of points at corresponding coordinate locations is zero.

We point out that this normalization criterion is more general and different from the one presented in [48]. There, the affine normalization step ensured a zero mean displacement and a mean transformation matrix that had determinant one.

4.2.7 Estimating Distributions

Smooth Histogramming

Histogramming is a widely used approach for approximating probability distributions. It keeps record of the frequency of occurrence of an observation within given fixed-width intervals (known as bins). Histograms, by definition, are not continuous functions. That can result in some undesirable discontinuities when, for example, computing transformation update terms. Thus numerous suggestions have been introduced about how to modify the *hard* assignments of the binning step to improve the smoothness property.

We implemented one such method that we refer to as the *smooth histogramming*

estimator. Each sample intensity is assigned to two as opposed to just one histogram bin. The value of the non-integer updates is assigned to the selected neighboring bins based upon the distance between the sample and the bin centers. This framework provided us with an accuracy similar to the one obtained with the Parzen Windowing approach (described below).

Parzen Windowing

The Parzen Windowing density estimator produces robust results and has the advantage that the derivatives of this estimator can be expressed in a more principled way than in the case of the histogram-based approach. If $[x_1, x_2, ..., x_N]$ are N samples of a random variable, then the Parzen window approximation of its probability distribution function is:

$$p(x) = \frac{1}{N} \sum_{i=1}^{N} K(x - x_i). \tag{4.17}$$

In our implementation, we selected the kernel function (K) to be a Gaussian

$$K(x) = G_\sigma(x) = \frac{1}{\sigma\sqrt{(2\pi)}} e^{-x^2/(2\sigma^2)}. \tag{4.18}$$

Ideally, the window size (the σ) of this kernel is computed online by optimizing a maximum likelihood objective function. We, however, fixed its value because empirical studies showed that changing the window size did not influence our results much, even when used with respect to different types of input images. In our experiments we set $\sigma = 2$.

4.2.8 Entropy Estimation

Although we experimented with several alternative methods, in our registration framework we chose to use the EMMA-style [77] iterative non-parametric approach for estimating entropies. This decision was chiefly influenced by the efficient gradient estimation facilitated by this framework. Our optimization framework also requires the explicit computation of the distribution of the sample intensities. Therefore, we implemented both widely-used smooth histogramming and a Parzen Windowing-style density estimators, where we know how to compute partial derivatives with respect to transformation components. The list and description of the other estimators that were considered are included in Sec. 2.5.

4.2.9 Stochastic Gradient-based Optimization

In the original framework of the congealing algorithm, a coordinate descent optimization was used to guide the minimization of the objective function. Such a non-gradient-based search was not feasible for our purposes. Because both the size and the number of image volumes are much larger in our proposed applications, memory allocation and computational speed are both of serious concern. Consequently,

we decided to apply a more efficient optimization strategy. Our choice favored an iterated stochastic gradient descent-based optimization as we have already achieved good results with it in the past [93, 77]. Using this framework, the conventional gradient descent optimization algorithm is combined with stochastic sampling and the EMMA-style[1] entropy estimator [84]. The stochastic sampling takes place in the spatial domain and not in the domain of the input volumes. Thus while all the inputs contribute to the estimation process, we propose a random selection of spatial coordinate locations in the display frame. The total sum of voxel-wise entropies corresponding to a particular alignment configuration is then approximated by using only a reduced set of samples. We write the modified objective function (approximating expectation with sample average) as:

$$f(\mathcal{U},\mathcal{T}) = -\frac{1}{N_A}\sum_{x_i \in A}\sum_{j=1}^{n} \log p^i(u_j(T_j(\mathbf{x_i}))), \tag{4.19}$$

where A now indicates the subset of randomly selected spatial coordinate locations, and N_A is equal to the cardinality of A. It is important to note that the samples in this reduced set are not fixed, but are re-generated at each iteration of the algorithm.

As the experiments show in Chapter 5, this modification enabled us to significantly speed up the search process. The reduction in the overall number of voxel locations visited per iteration provides a tremendous increase in computation speed. Frequently, we manage to carry out successful registration when only examining .1% of the data samples at each iteration.

4.2.10 The Gradient-based Update Computations

According to the EMMA-style non-parametric entropy estimator using Parzen Windowing for density estimation, the sample entropy H^* of random variable Z can be expressed as follows:

$$\begin{aligned} H(Z) \approx H^*(z) &= -\frac{1}{N_B}\sum_{z_i \in B} \ln P^*(z_i) &(4.20)\\ &= -\frac{1}{N_B}\sum_{z_i \in B} \ln \frac{1}{N_A}\sum_{z_j \in A} G_{\Psi_z}(z_i - z_j), &(4.21)\end{aligned}$$

where A and B refer to data samples of the random variable; N_A and N_B refer to the number of elements in those two samples respectively, z_i and z_j represent two instances of the random variable, and G_Ψ is the Gaussian kernel of zero mean and Ψ covariance matrix. The derivative of the sample entropy term, where the random

[1]EMMA is a random but pronounceable subset of the letters in the words "Empirical entropy manipulation and analysis"[77].

variable Z is dependent on the parameter T, can be computed via [77]:

$$\frac{d}{dT}H^*(Z) = \frac{1}{N_B}\sum_{z_i \in B}\sum_{z_j \in A} W_z(z_i, z_j)(z_i - z_j)^T \Psi_z^{-1}\frac{d}{dT}(z_i - z_j), \tag{4.22}$$

where we define $W_z(z_i, z_j) \equiv \frac{G_{\Psi_z}(z_i - z_j)}{\sum_{x_k \in} G_{\Psi_z}(z_i - z_k)}$. We rely on this formulation when deriving our optimization updates.

The objective function of the congealing algorithm is the sum of voxel-wise entropy measures. Using the modified objective function from Eq.(4.19) and the sample entropy definition in Eq. (4.21), the former can be expressed as:

$$\begin{aligned} f(\mathcal{U}, \mathcal{T}) &= \sum_{x_i \in A} H(\hat{p}(\mathcal{U}(\mathcal{T}(\mathbf{x_i})))) \\ &\approx \sum_{x_i \in A} H^*(\hat{p}(\mathcal{U}(\mathcal{T}(\mathbf{x_i})))) \\ &= -\frac{1}{N_A}\sum_{x_i \in A}\sum_{j=1}^{n} \log p^i(u_j(T_j(\mathbf{x_i}))) \\ &\approx -\frac{1}{N_A}\sum_{x_i \in A}\sum_{j=1}^{n} \log \frac{1}{N_A}\sum_{k=1}^{n} G_\Psi(u_j(T_j(\mathbf{x_i})) - u_k(T_k(\mathbf{x_i}))), \end{aligned} \tag{4.23}$$

where again $\mathbf{x_i}$ indicates a coordinate location in R^3 and A is the set of samples randomly selected in the stochastic gradient-based optimization framework.

In order to find the best alignment for the input data sets, this objective function needs to be minimized with respect to the collection of transformations $\mathcal{T} = \{T_1, T_2, ..., T_n\}$. At each iteration of the registration process we update the current estimate of these transformations according to Eq. (4.24).

$$\mathcal{T}_{\text{next}} \leftarrow \mathcal{T} + \lambda \frac{d}{d\mathcal{T}} f(\mathcal{U}, \mathcal{T}), \tag{4.24}$$

where the λ term, called the *learning rate*, controls the maximum allowable step size throughout the optimization.

Thus we are interested in computing the partial derivatives of the objective function with respect to each of the transformation components.

$$\frac{d}{d\mathcal{T}} f(\mathcal{U}, \mathcal{T}) \approx \sum_{x_i \in A} \frac{\partial}{\partial \mathcal{T}} H^*(\mathcal{U}(\mathcal{T}(x_i))). \tag{4.25}$$

In order to simplify the notation in the upcoming derivation section, we define $y_i \equiv \mathcal{U}(\mathcal{T}(x_i))$ to represent all intensity samples from the inputs collected at a particular

coordinate location x_i. More precisely,

$$y_i = \{u_1(T_1(x_i)); u_2(T_2(x_i)); ...; u_n(T_n(x_i))\}. \tag{4.26}$$

Thus for the jth data volume, $(1 \leq j \leq n)$, $y_{ij} \equiv u_j(\mathcal{T}_j(x_i))$.

Again, if N_A signifies the total number of voxels selected in our optimization framework, and n is the number of input volumes, using the definition in Eq.(4.22):

$$\begin{aligned}
\frac{d}{d\mathcal{T}} f(\mathcal{U}, \mathcal{T}) &\approx \sum_{x_i \in A} \frac{\partial}{\partial \mathcal{T}} H^*(\mathcal{U}(\mathcal{T}(x_i))) && (4.27)\\
&\approx \sum_{x_i \in A} \frac{1}{N_A} \sum_{j=1}^{n} \sum_{k=1}^{n} W_{y_i}(y_{ij}, y_{ik})(y_{ij} - y_{ik})^T \Psi_y^{-1} \frac{d}{dT}(y_{ij} - y_{ik}) && (4.28)\\
&= \frac{1}{N_A} \sum_{x_i \in A} \sum_{j=1}^{n} \sum_{k=1}^{n} W_{y_i}(y_{ij}, y_{ik})(y_{ij} - y_{ik})^T \Psi_y^{-1} \frac{d}{dT}(y_{ij} - y_{ik}), && (4.29)
\end{aligned}$$

where we introduce W to simplify the notation and we also make the assumption that the covariance matrices are diagonal:

$$W_{y_i}(y_{ij}, y_{ik}) \equiv \frac{G_{\Psi_y}(y_{ij} - y_{ik})}{\sum_{l=1}^{n} G_{\Psi_y}(y_{ij} - y_{il})} \quad \text{and} \quad \Psi_y = \text{DIAG}(\sigma_{y_j}^2, \sigma_{y_l}^2). \tag{4.30}$$

To compute the optimization updates at a given iteration, we need to compute the updates for each individual input data. Thus for each $l \in [1, 2, ..., n]$:

$$\begin{aligned}
&\frac{d}{d\mathcal{T}_l} f(\mathcal{U}, \mathcal{T}) = \\
&= \frac{1}{n} \sum_{x_i \in A} \sum_{j=1}^{n} \sum_{k=1}^{n} W_{y_i}(y_{ij}, y_{ik})(y_{ij} - y_{ik})^T \Psi^{-1} \frac{d}{d\mathcal{T}_l}(y_{ij} - y_{ik}) \\
&= \frac{1}{n} \sum_{x_i \in A} \left[\sum_{k=1}^{n} W_{y_i}(y_{il}, y_{ik})(y_{il} - y_{ik})^T \Psi^{-1} \frac{d}{d\mathcal{T}_l}(y_{il}) - \sum_{j=1}^{n} W_{y_i}(y_{ij}, y_{il})(y_{ij} - y_{il})^T \Psi^{-1} \frac{d}{d\mathcal{T}_l}(y_{il}) \right] \\
&= \frac{1}{n} \sum_{x_i \in A} \left[\sum_{z=1}^{n} W_{y_i}(y_{il}, y_{iz})(y_{il} - y_{iz})^T \Psi^{-1} \frac{d}{d\mathcal{T}_l}(y_{il}) - W_{y_i}(y_{iz}, y_{il})(y_{iz} - y_{il})^T \Psi^{-1} \frac{d}{d\mathcal{T}_l}(y_{il}) \right].
\end{aligned}$$

Because the Gaussian kernels are zero mean, the parameters of W can be inter-

changed, $W_{y_i}(y_{ij}, y_{il}) = W_{y_i}(y_{il}, y_{ij})$, and we write:

$$\begin{aligned}
\frac{d}{d\mathcal{T}_l} f(\mathcal{U}, \mathcal{T}) &\approx \frac{1}{n} \sum_{x_i \in A} \left[\sum_{z=1}^{n} \left[(y_{il} - y_{iz})^T - (y_{iz} - y_{il})^T \right] W_{y_i}(y_{iz}, y_{il}) \Psi^{-1} \frac{d}{d\mathcal{T}_l} y_{il} \right] \\
&\approx \frac{1}{n} \sum_{x_i \in A} \sum_{z=1}^{n} \left[2(y_{il} - y_{iz})^T W_{y_i}(y_{iz}, y_{il}) \Psi^{-1} \frac{d}{d\mathcal{T}_l} y_{il} \right] \qquad (4.31) \\
&\approx \frac{2}{n} \sum_{x_i \in A} \sum_{z=1}^{n} (y_{il} - y_{iz})^T W_{y_i}(y_{iz}, y_{il}) \Psi^{-1} \frac{d}{d\mathcal{T}_l} y_{il}. \qquad (4.32)
\end{aligned}$$

The fourth term in Eq.(4.32) corresponds to the partial derivative of the intensity values with respect to the transformation components. According to the matrix notation that we use, that term can be written as:

$$\frac{d}{d\mathcal{T}_l} y_{il} = \frac{d}{d\mathcal{T}_l} u_l(\mathcal{T}_l(x_i)) = \begin{cases} \frac{d}{d\mathcal{D}_l} u_l(\mathcal{T}_l(x_i)) & : \quad \nabla u_l(\mathcal{T}_l(x_i)) \\ \frac{d}{d\mathcal{K}_l} u_l(\mathcal{T}_l(x_i)) & : \quad \nabla u_l(\mathcal{T}_l(x_i)) {x_i}^T \end{cases} \qquad (4.33)$$

These derivations are also useful when defining optimization using free-form deformation fields. There, specifically in the case of a dense deformation map, a local displacement vector is defined at each grid or voxel location. Thus the update term with respect to the displacement field D in Eq. (4.33) will be used.

4.2.11 Initialization

The algorithm is very robust with respect to initialization. As our experiments in Chapter 5 show, no pre-alignment or careful initialization was required before the registration process. Many truncated and bad quality images have been processed and their presence did not offset the alignment results. In addition, the registration of these outlier images has become more robust with the increasing number of input volumes. Our hypothesis is, that the more the image population grows, the less the effect of outliers become on the registration results.

4.2.12 Alignment of New Observations to the Group

Once the initial set of brain volumes has been registered, aligning a new data sample to the statistical model is simple [48]. The group-wise registration framework not only results in a low total entropy "distribution image" that represents the shape and its residual variation that can not be further reduced, but also a set of final transformations that relate each input to the others. Thus, constructing a probability distribution over the transformation domain could provide a good initialization to the registration of a new data sample to the registered collection (or the distribution image).

4.3 Summary

In this chapter, we described the group-wise registration problem as a generalization to the pair-wise alignment problem. We positioned the objective function of several existing methods into our extended unified information theoretic framework and also introduced the congealing framework, which we selected as the basis for our new and efficient group-wise registration framework. In the following chapter, we demonstrate the performance of that new registration framework via a wide selection of experiments. We use numerous data sets and validate the accuracy and the repeatability property of the algorithm.

Chapter 5

Group-wise Registration Experiments and Validation

In this chapter, we demonstrate the performance characteristics of our group-wise registration algorithm, stochastic congealing*, through a set of carefully defined validation experiments. We present the advantages of the method and also quantitatively evaluate the accuracy of our results.*

5.1 Experiments Using Medical MRI Data Sets

In all the upcoming sections except for the last one of this chapter, our registration experiments are optimized with respect to affine transformations. In the last section, we extend our formulation to also use higher dimensional free-form deformations. The timing results for individual experiments were all measured on Intel(R) Xeon(TM) 3.2GHz machines.

5.1.1 Visualization

The visual representation of the registration results is not a trivial task. Given the high number and the high resolution of our input data volumes, it is not intuitive what type of display would best facilitate the evaluation of the registration outcomes. Below and in Chapter 6, we use two different methods to show how well the data volumes align after the registration algorithm. With one method, we compute the mean volume from all the input data (in corresponding position) and display three orthogonal slices of that, and with the other, we display one particular corresponding slice (usually the central coronal one) from all the input volumes. We do not claim that these visualization methods are optimal, but they convey complimentary information and allow for visual appreciation of the registration performance.

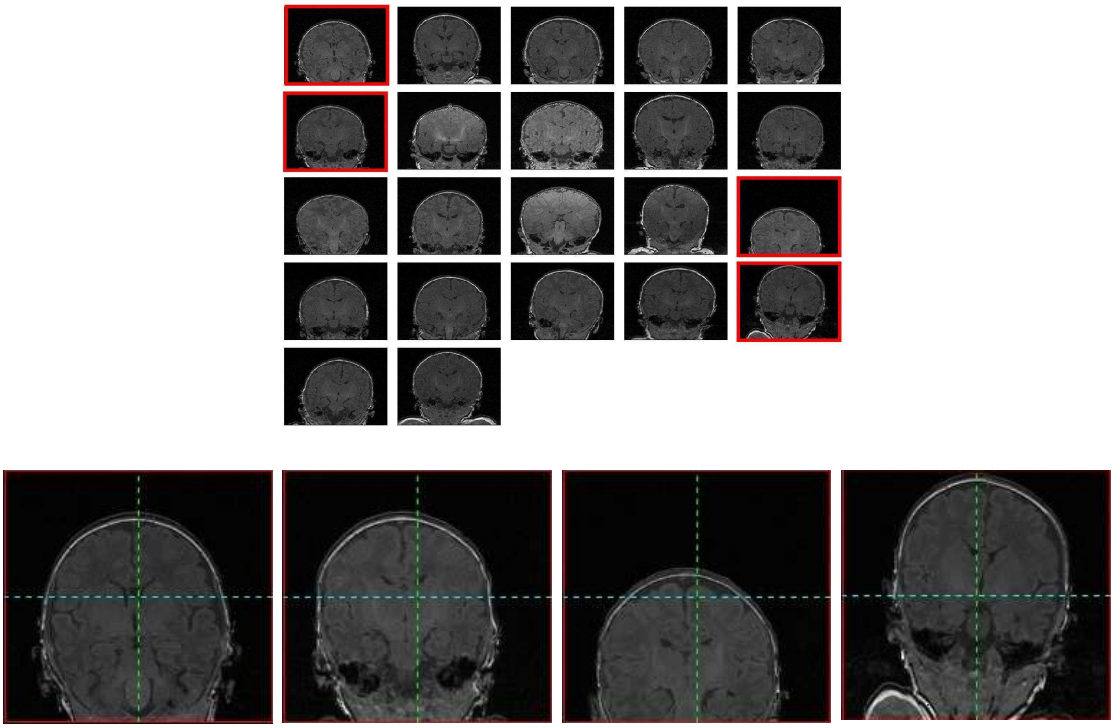


Figure 5-1: The baby brain data set of 22 T1-weighted MRI volumes. Central coronal slices of the input images were obtained at the initial, misaligned position. Four of the slices (framed with a red box in the top image) are enlarged in order to better demonstrate the within group differences.

5.1.2 Data Set Description

We start the discussion of this experimental section by demonstrating the registration results on two challenging data sets. One is a collection of T1-weighted head data volumes taken of babies and the other is collection of T1-weighted MRI data volumes of adult subjects. The image processing challenges for these data sets not only originate from the size of the data sets, but also from the varying quality of the data samples. Many of our data sets are corrupted by noise from the imaging device or by motion artifacts due to non-cooperative subjects (e.g., babies). We point out these and other complexities in more details corresponding to particular experiments.

The Baby Brain Scans

The first medical data set was provided to us by the research group Simon K. Warfield, Ph.D. (Computational Radiology Laboratory, Department of Radiology, Brigham and Women's Hospital). It consists of twenty-two baby brain volumes of T1-weighted MRI. In size, each brain volume is 176 by 186 by 110 voxels, with each voxel measuring 1.0 by 1.0 by 2.0 millimeters. The central coronal slices belonging to that uni-modal population are displayed in Fig. 5-1. In order to better appreciate the scale of the within-group differences, we enlarged the images of four of the subjects (framed with

a red box in the top image). It is clear, even from these two-dimensional images, that the imaged anatomies are highly variable in size and shape, some of them are partially cropped (third enlarged image at the bottom) and they also demonstrate image intensity differences and bias field corruption. One can also notice that some of the structures are not yet fully developed (the folding patterns are highly variable, for example), and frequently motion artifacts also degrade the image quality. (It is extremely challenging to arrange for the infant subjects to remain immobile for the full duration of imaging.)

Adult Brain Scans

The second set of data was obtained from the group of Martha Shenton, Ph.D. (Department of Psychiatry, Harvard Medical School) through the NAMIC[1] consortium. The collection consists of twenty-eight T1-weighted MRI acquisitions of adult brains. These volumes are 256 by 256 by 124 voxels, with each voxel measuring 0.9375 by 0.9375 by 1.5 millimeters. The central coronal slices of these volumes are displayed in Fig. 5-2. Again, we enlarged the images of four of the subjects for an easier evaluation of the results.

The acquisitions in the adult data set are of higher resolution, but the volumes are corrupted by some non-uniform bias. Our experimental results show in the next section that the algorithm can properly deal with these artifacts.

The Experiments

The corresponding experiments were run at two different resolution levels in sequence. Most of the time, the final convergence result was almost completely achieved on the lower resolution level and it was only a very fine refinement that was needed on the highest resolution level. In both of these experiments we only had to select between .05-.1% of the total voxels per iteration, and no more than 300 iterations were sufficient to recover the aligning affine transformations. The total running time, respectively for the two experiments, was 340 and 1209 seconds, which considering the size and the number of data volumes are very promising.

The results of the baby brain experiments are displayed in Fig. 5-3- 5-5. Figure 5-3 displays the central coronal slice of each of the input volumes (a) before and (b) after the alignment. Figure 5-4 demonstrates the same results via the pre-selected and enlarged set of four subjects. The top row presents the images before and the bottom row after registration. Three orthogonal slices of the mean volumes computed before and after the experiments are shown in Fig. 5-5. We can establish that following the group-wise alignment, the data volumes properly line up and the mean volumes have

[1]Information about the National Alliance for Medical Image Computing (NAMIC) and the National Centers for Biomedical Computing can be obtained from http://nihroadmap.nih.gov/bioinformatics

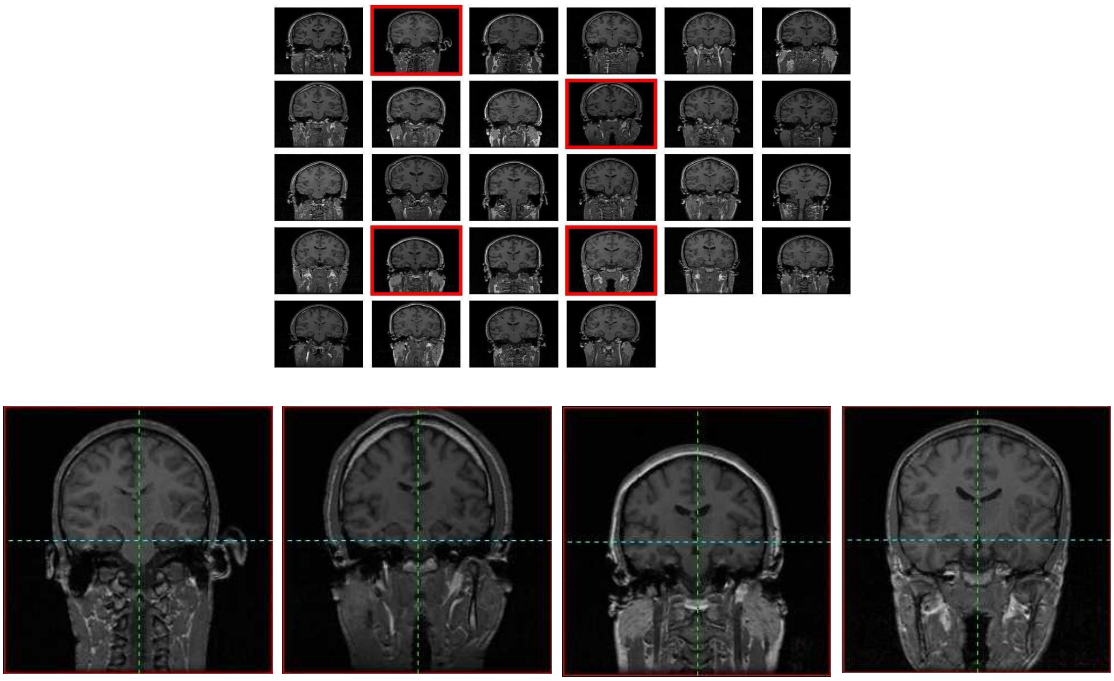
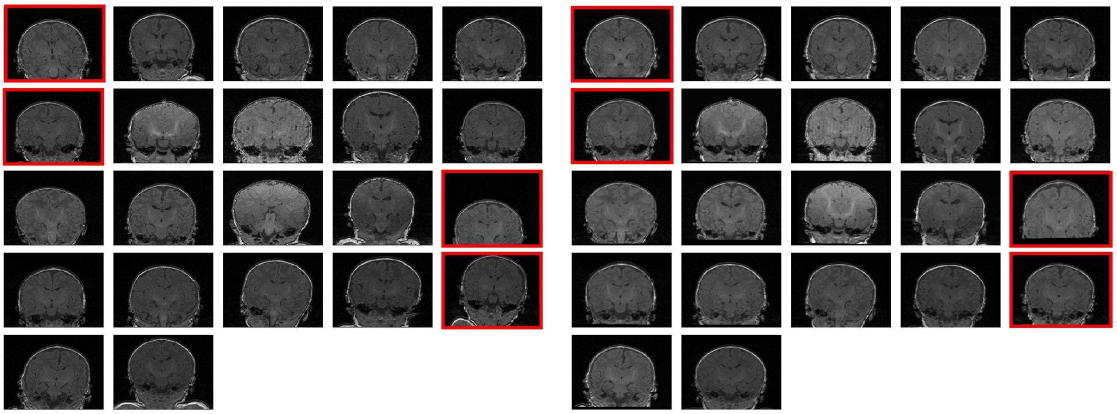


Figure 5-2: The adult brain data set of 28 T1-weighted MRI volumes. Central coronal slices of the input images were obtained at the initial, misaligned position. Four of the slices (framed with a red box in the top image) are enlarged in order to better demonstrate the within group differences.

clean and sharp boundaries.

The outcome of the adult brain experiments can be seen in Fig. 5-6 - 5-8. Figure 5-6 (a) displays the central coronal slice of each of the input volumes before and (b) after the alignment. Figure 5-7 displays the same results with the pre-selected and enlarged images. Three orthogonal views of the mean volumes computed before and after registration from these data sets is displayed in Fig. 5-8. After the initial mismatch indicated by the blurriness in (a), the data volumes properly align producing a mean volume with clear and sharp boundaries (b).



(a) Before congealing (b) After congealing

Figure 5-3: The baby brain data set of 22 MRI volumes. Central coronal slices of selected input images were obtained (a) before and (b) after the stochastic congealing process.

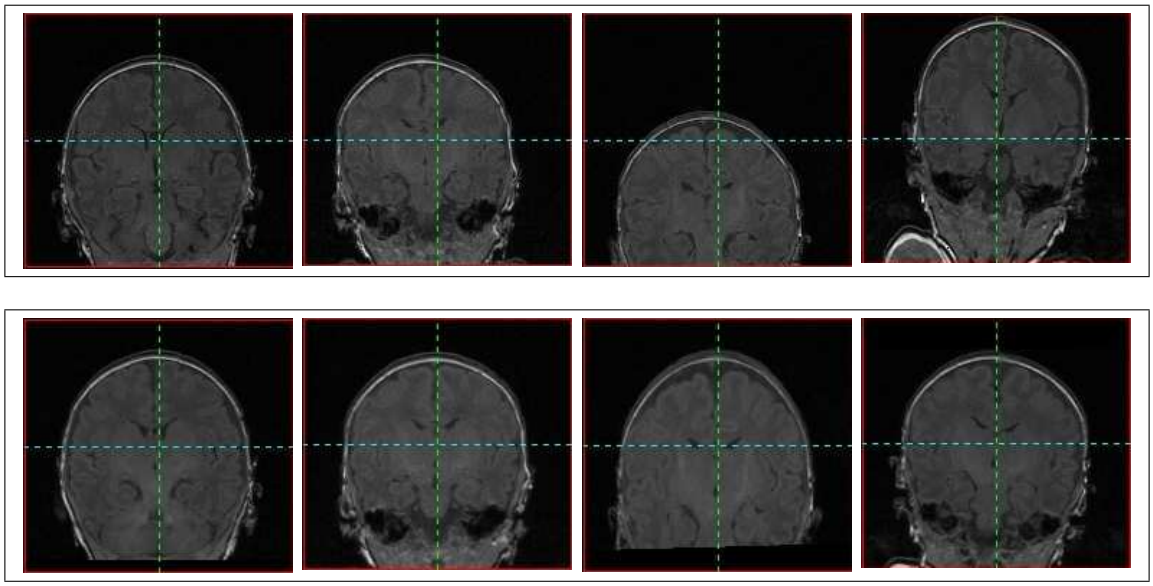
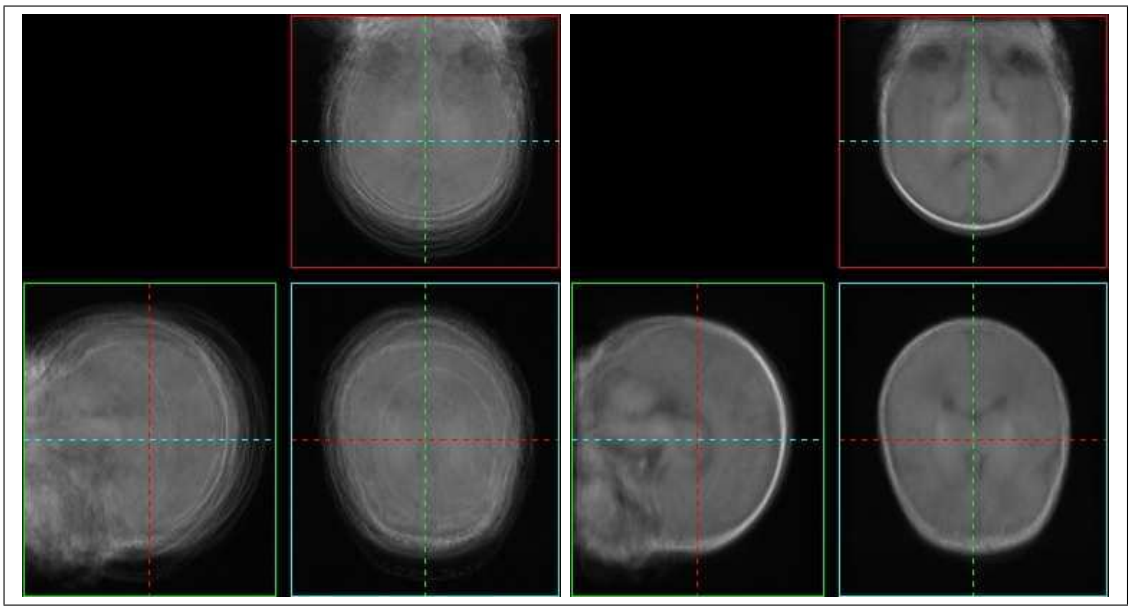


Figure 5-4: The baby brain data set of 22 T1-weighted MRI volumes. Central coronal slices of the input images were obtained (top row) before and (bottom row) after the stochastic congealing process.



(a) Before congealing (b) After congealing

Figure 5-5: Three orthogonal views of the mean volume created from the baby brain data set: (a) before and (b) after the stochastic congealing process.

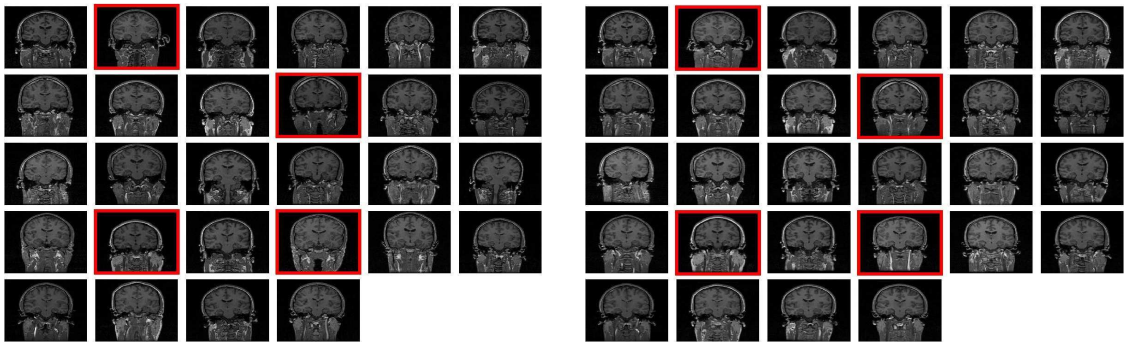


Figure 5-6: The adult brain data set of 28 MRI volumes. Central coronal slices of the input images (a) before and (b) after the stochastic congealing process.

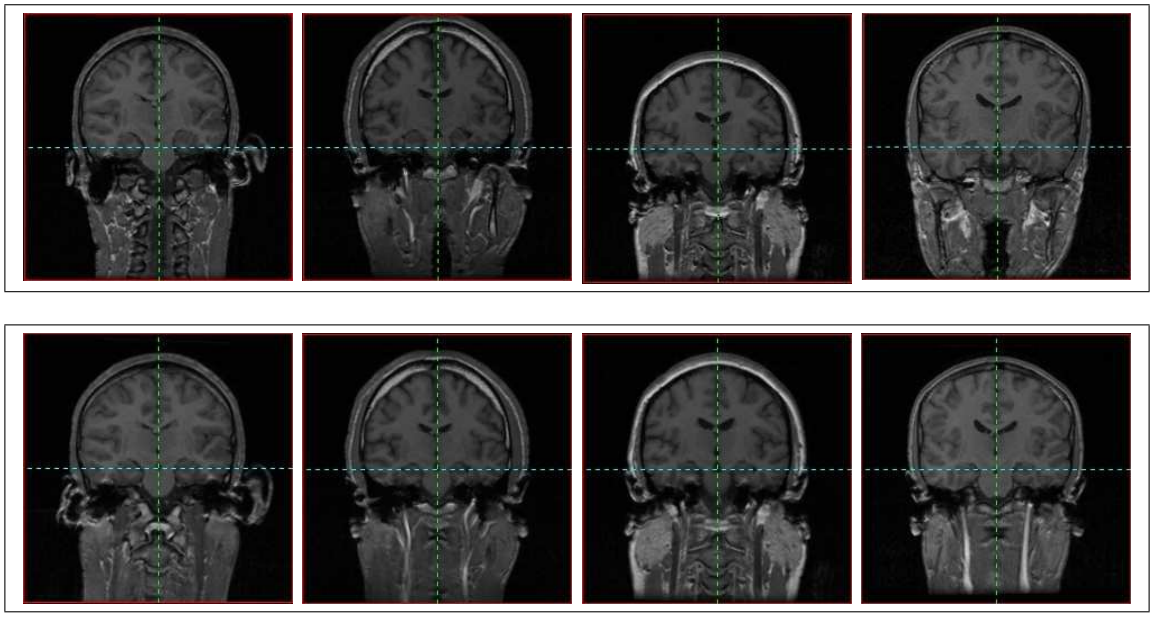
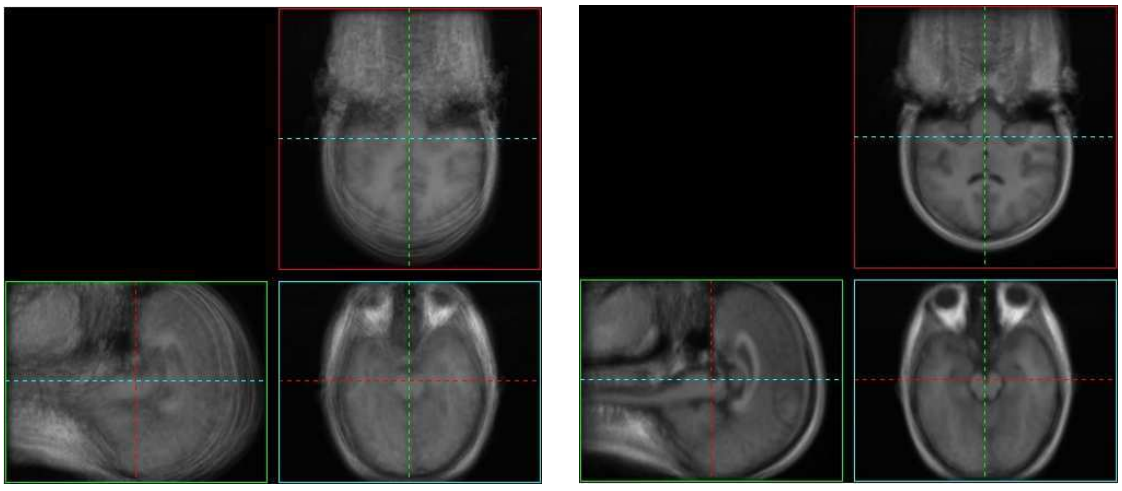


Figure 5-7: The adult brain data set of 28 T1-weighted MRI volumes. Central coronal slices of the input images were obtained (top row) before and (bottom row) after the stochastic congealing process.



(a) Before congealing (b) After congealing

Figure 5-8: Three orthogonal views of the mean volume created from the adult brain data population of 22 images: (a) before and (b) after the stochastic congealing process.

5.1.3 Large Data Set Registration

As described earlier, one of the main advantages of our stochastic congealing group-wise registration framework is that it allows for the alignment of very large data sets. While some other methods in the field have to face increasing difficulties as the number of input images gets larger (often the number of required samples grows exponentially for a reliable estimate of joint entropy whose dimensionality is linearly growing with respect to the number of data volumes) for our framework the entropy estimation procedure remains one-dimensional. As more input volumes result in more samples, our method indeed favors larger populations. The more input volumes there are, the more reliable its entropy estimation becomes. In this section we present the largest data set that we processed and demonstrate the feasibility of an efficient registration even in the case of 127 input data volumes. We believe that our group was the first to report the simultaneous registration run on such a large collection of input 3D volumes.

Our *large* data collection consists of 127 adult brain volumes of T1-weighted MRI images. These volumes are 256 by 256 by 124 voxels, with each voxel measuring 0.9375 by 0.9375 by 1.5 millimeters. Fig. 5-9 displays the central coronal slices extracted from each member of that population. Even though the individual images are a bit too small for all the details to be clearly visible, the figure still represents some of the variability in the data set.

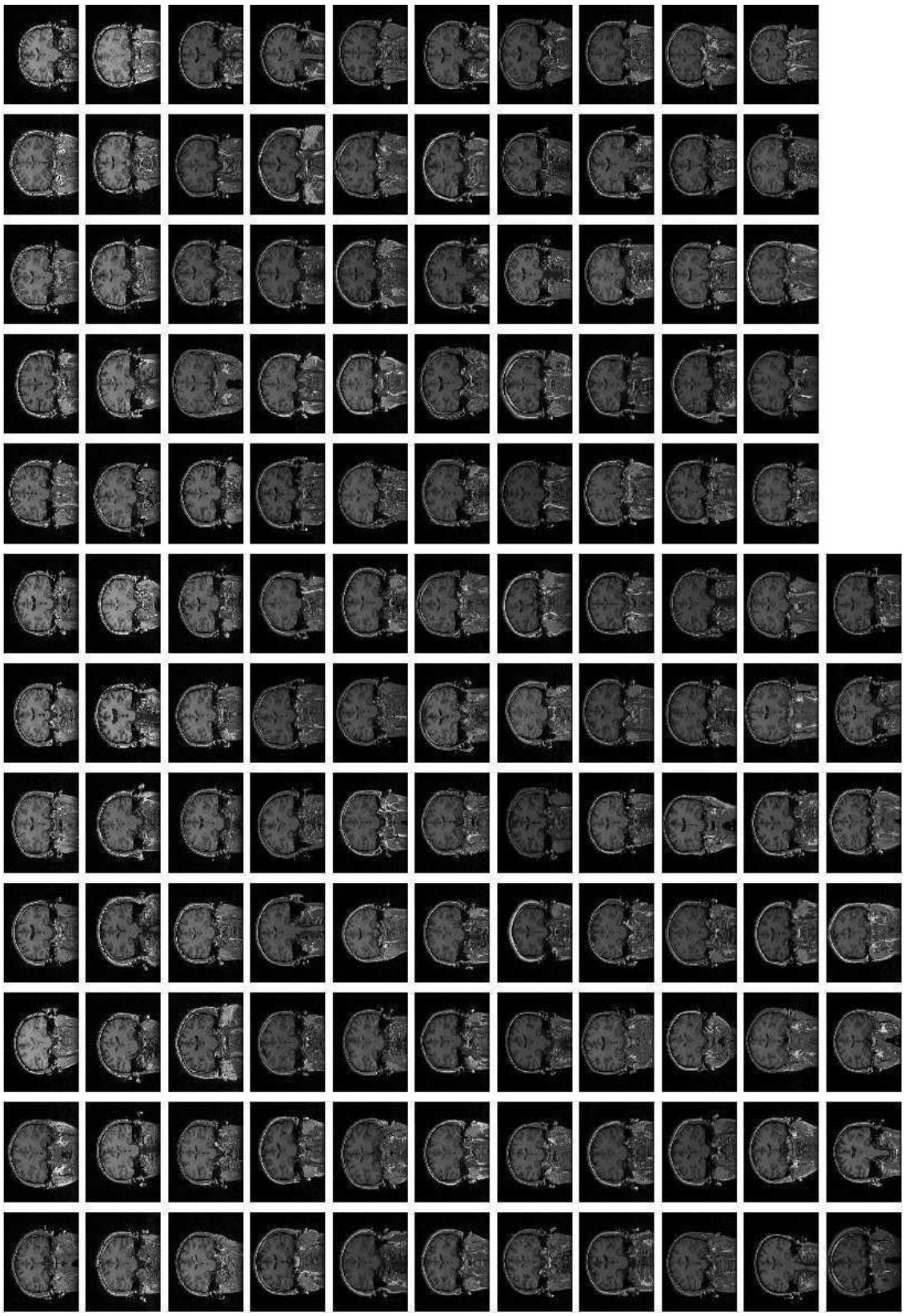
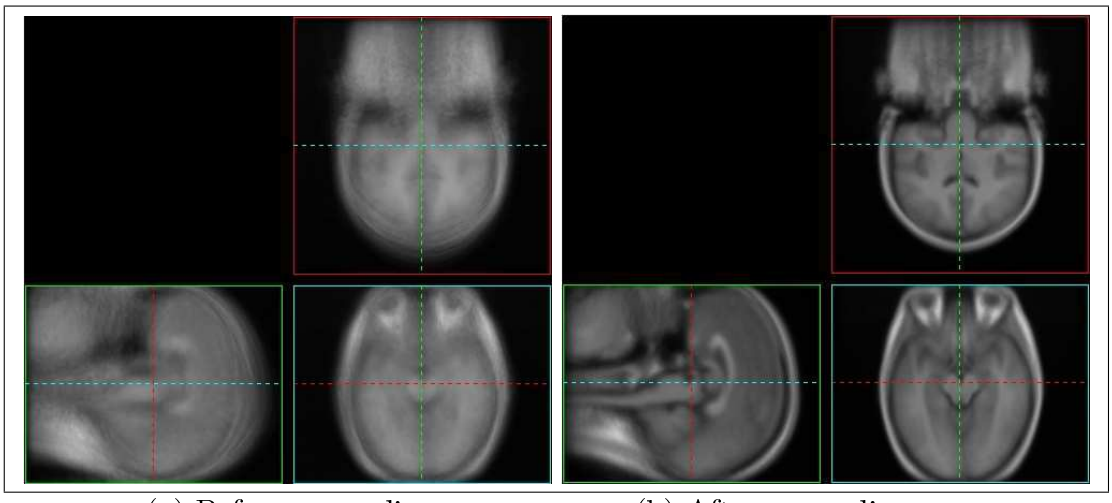


Figure 5-9: The adult brain data set of 127 MRI volumes. Central coronal slices of the unaligned input images.

The affine experiments on the 127 scans were executed on three different resolution levels (where the volumes were smoothed and downsampled to (32 by 32 by 31), (64 by 64 by 62) and (128 by 128 by 124) voxels). The largest offset was obtained on the lowest level and then refinement was computed on the higher hierarchy levels. In our experiments we only had to select between 800 - 1500 samples, which constitutes just .05-2.5% of the total voxels per iteration, and no more than 250 iterations were sufficient in order to obtain high quality alignment. The total running time for the experiment was approximately six hours.

The results of the experiments are displayed in Figure 5-10. It shows three orthogonal slices of the mean volumes computed (a) before and (b) after the experiments. We can again establish, qualitatively, that the the mean volumes have clean and sharp boundaries after the group-wise alignment.



(a) Before congealing (b) After congealing

Figure 5-10: Three orthogonal views of the mean volume created from the 127 MRI volumes of our large data set: (a) before and (b) after the stochastic congealing process.

5.1.4 Minimum Number of Data Sets

In the preceding experiments, we demonstrated the attractive properties of our registration framework on a large set of data. There is another question that arises: what is the smallest number of images that our registration framework can handle? Although experimentally we have successfully registered a data set of five images, the general answer is slightly more complex. Theoretically, we need as many input data sets as the number of samples that are necessary to provide a sufficiently robust entropy estimate from the observed intensity samples. In general, the fewer the number of the data samples, the more fragile our entropy estimator becomes. However, the estimation also depends on the quality of the input images. We expect to achieve successful registration with even just a small set of high quality and noise-free images, but when working with lower quality data sets the number of input data would need to be increased.

One simple modification that could potentially increase the robustness of the stochastic congealing framework with respect to the minimum number of input images is the implementation of the sum of *local neighborhood* entropies [34]. Instead of estimating entropies corresponding to a single voxel location, we could draw intensity samples from the local neighborhood of voxels thus increasing the number of samples based upon which we build our entropy estimator. The detailed evaluation of this idea is yet to be completed.

5.2 Multi-modal Image Data Set Registration

As we mentioned in our literature review in Chapter 2, the majority of the currently used group-wise registration algorithms are only suitable for uni-modal registration tasks. That is because their objective functions are not robust enough to model / handle different intensity profiles characterizing the imaged anatomy. Although the set of information theoretic methods that we analyzed in Chapter 4 do have the capability of accommodating a wider variety of input image profiles, they are limited in the number of input data sets that they can operate on. Their applicability is limited by relying on increasing dimensional density- and entropy estimation tasks.

It is feasible to accommodate multi-modal data sets in the congealing framework. The key is to make sure that the identicality assumption in the imaging domain remains valid. That is to say that samples collected from the same spatial coordinate location from the input data sets can be explained by the same distribution, satisfying the *iid* condition. We found that given different types of MRI images the stochastic congealing registration still works successfully even with single 1D entropy estimations. For the registration of significantly different modalities, such as CT, MRI and PET for example, higher dimensional entropy estimations are necessary as the identicality assumption otherwise does not hold. In the following section we provide registration results with respect to an MRI data set containing T1-, T2- and

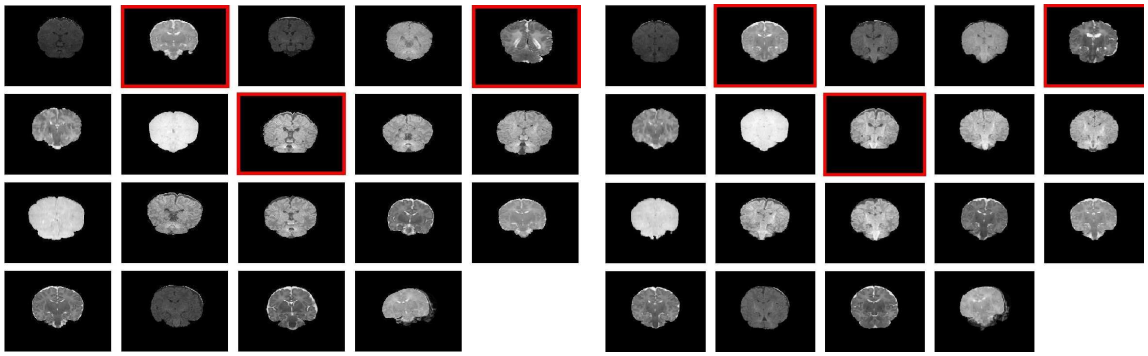
PD-weighted acquisitions.

5.2.1 Pre-term Baby Brain Scans

The collection of pre-term baby head scans contains acquisitions for twenty different subjects with three different modalities. The volumes are 256 by 256 by 110 voxels, with each voxel measuring 0.703125 by 0.703125 by 1.5 millimeters and the three image modalities are: T1-, T2- and PD-weighted MRI. From the total of 60 images, we randomly selected nineteen corresponding to ten T1-, five T2- and four PD-weighted images. In this experiment the input volumes are *skull-stripped*. This means that we applied a mask of the intra-cranial cavity in order to segment out the within skull brain regions (gray matter, white matter and cerebro-spinal fluid).

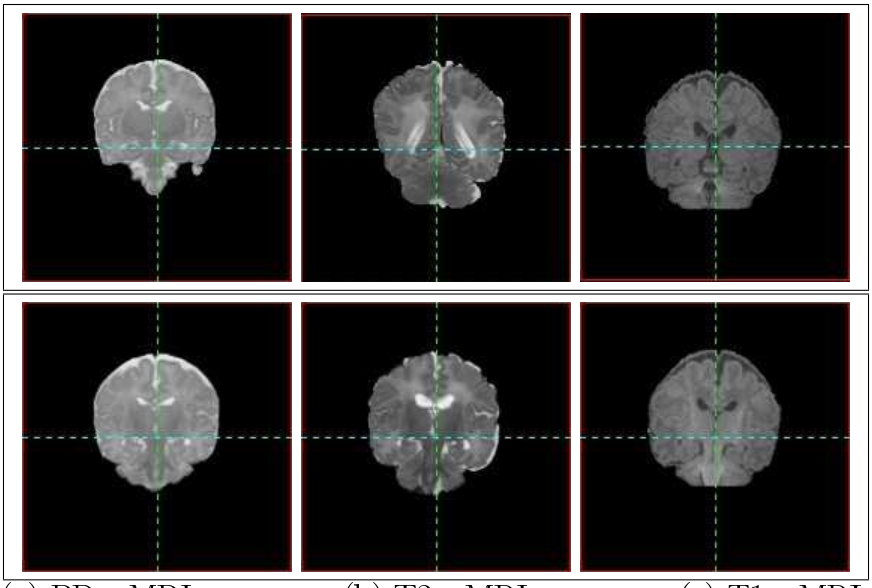
5.2.2 Experiments

In Figures 5-11-5-13, we demonstrate the registration results of stochastic congealing corresponding to the multi-modal pre-term data set. Even though the input data sets are of very different quality and many of them are even cropped, our registration managed to align them with good accuracy. In Figure 5-11, we show the set of unaligned and aligned input images through their corresponding coronal slices. Three pre-selected images are enlarged and displayed in Fig. 5-12. Lastly, besides the 2D slices, we present three orthogonal views of the mean volumes created before and after alignment in Fig. 5-13. It is easy to establish that the alignment resulted in clearer and sharper boundaries.



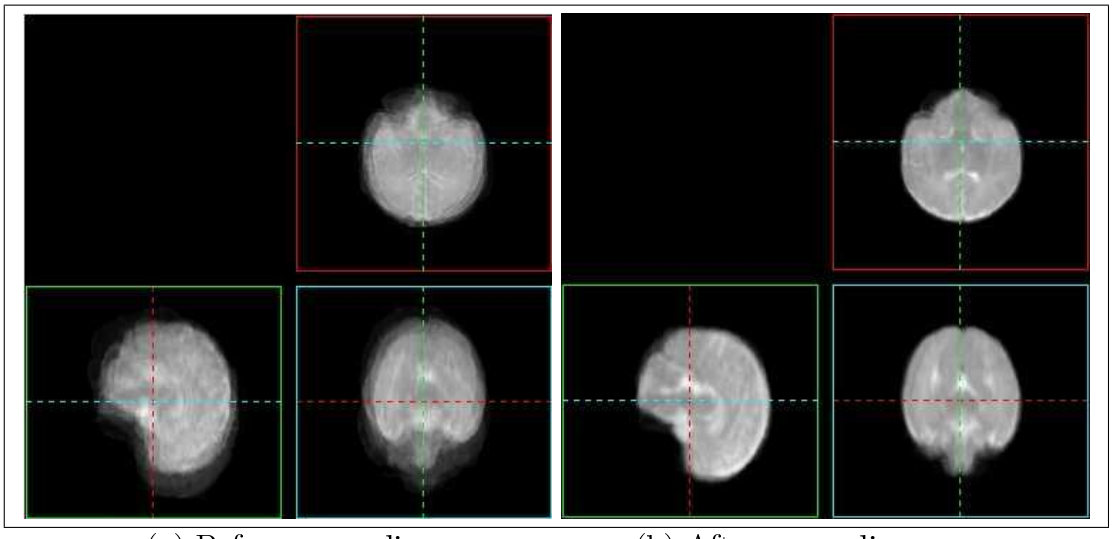
(a) Before congealing (b) After congealing

Figure 5-11: Central coronal slices of the multi-modal data set of three different types of baby brain acquisitions (a) before and (b) after the stochastic congealing. The image modalities are: PD-, T1- and T2-weighted images.



(a) PDw MRI (b) T2w MRI (c) T1w MRI

Figure 5-12: The multi-modal pre-term baby head data set consisting of 20 MRI volumes. Corresponding coronal slices of the input volumes are shown (top row) before and (bottom row) after the stochastic congealing process.



(a) Before congealing (b) After congealing

Figure 5-13: Three orthogonal views of the mean volumes created from the multi-modal baby data set (a) before and (b) after the stochastic congealing process.

5.3 Validation Experiments

As mentioned and demonstrated in the above sections, the results of the congealing process are qualitatively good and appealing. We pointed out that the mean volumes constructed after registration have much sharper boundaries than prior to alignment and examining the central coronal slices also suggests that good quality alignment is established.

In the current section, we provide a quantitative analysis in order to evaluate the accuracy of our framework. As obtaining ground truth registration parameters was not feasible in most of our data sets, we designed a set of experiments to describe different error characteristics of the alignment results. Using both synthetic and real collections of 3D data sets, we evaluate both the accuracy and the repeatability of the framework. Finally, we compare the quality of the mean volume representation produced by stochastic congealing with that created by another currently used method.

5.3.1 Synthetic Example

For our first analysis, we created a synthetic data set. After selecting one particular MRI volume from a group of adult brain acquisitions, we applied random affine transformations to it and thus created forty volumes. The magnitude of these transformations varied between $+/-10$ degrees for rotation, $+/-10$ mm for displacement, between $[.85, 1.15]$ factors for scaling and between $+/-.1$ factors of shearing. All the input volumes were 124 by 256 by 256 voxels, with each voxel measuring .9375 by .9375 by 1.5 mm. A simple figure illustrating this process is presented in Fig. 5-14. We refer to the notation indicated there in the upcoming analysis.

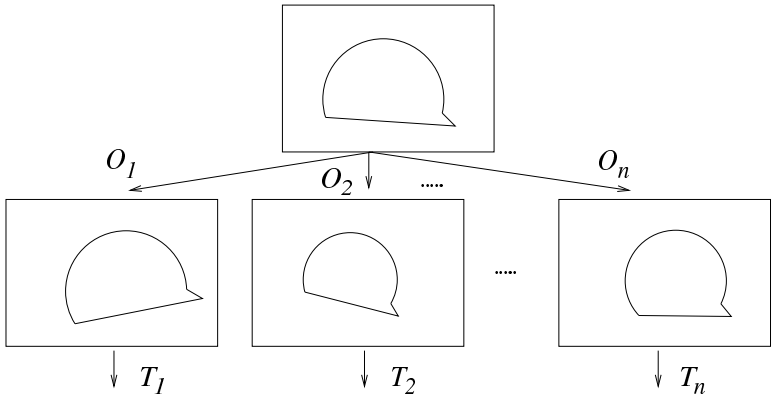


Figure 5-14: A schematic figure illustrating how the synthetic data set is created. We refer to the indicated notation in Sec.5.3.1.

Overlap Analysis

The first set of quantitative results that we present were computed using segmentation results. More precisely, given a set of classification labels identified in the original im-

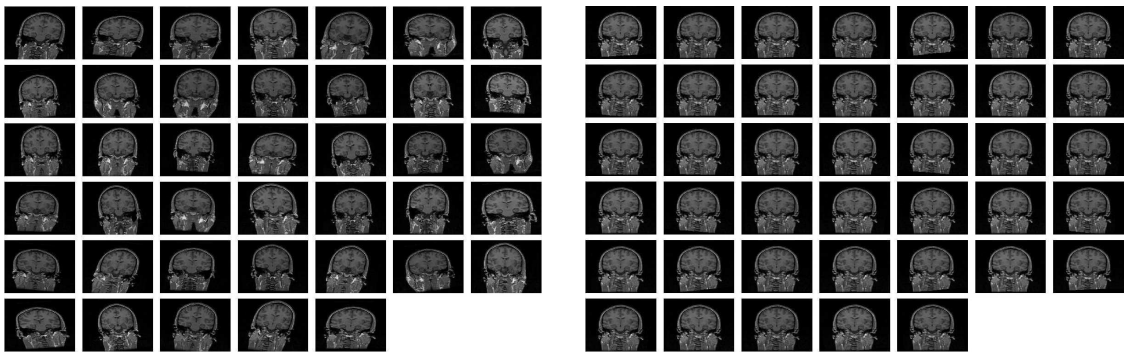
age, we measured label agreement in the first randomly offset and then aligned data. Initially, we had access to the segmentation of the Intra-Cranial Cavity (ICC) which is composed of the grey matter (GM), white matter(WM) and cerebro-spinal fluid (CSF) in the brain. These we offset with the same set of random transformations that we applied to the gray-scale images. After the congealing alignment was completed, we also applied the resulting transformations to them. Then we computed an overlap measure between the initial and the aligned ICC binary maps.

We selected an overlap indicator that could be easily generalized to higher number of inputs. The measure we chose (where A_i indicate binary variables) was

$$f_{\text{overlap}}(A_1, A_2) = \frac{|A_1 \bigcap A_2|}{min(|A_1|, |A_2|)}, \tag{5.1}$$

or in words, the ratio of the area of intersection over the minimum of the input areas.

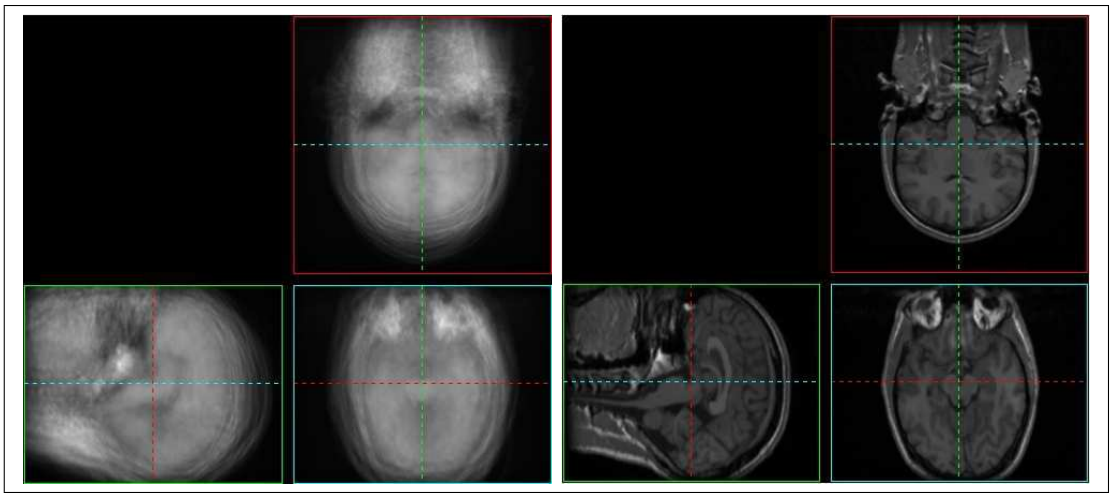
The twelve parameters of the affine transformations were nicely recovered after running our algorithm on two levels of the hierarchy. The number of samples used was small, only .05% of the total number of voxels. Fewer than 400 iterations were necessary to achieve convergence. The total running time was 2964 seconds. The results of these experiments can be seen in Fig. 5-15 and 5-16. The former illustrates the central coronal slices of each of the input volumes before and after the alignment and the latter displays the mean volumes computed before and after the congealing process.



(a) Before congealing (b) After congealing

Figure 5-15: Synthetic data set of 40 MRI volumes. Central coronal slices of the input images (a) before and (b) after the stochastic congealing process.

The overlap score also indicates great improvement in the alignment. From the original unaligned scenario of overlap percentage 51.08%, the metric increased to 95.98%. While the later score is not perfect, it indicates a very high quality alignment. We believe that there are two reasons for why we did not achieve 100% accuracy. First, the overlap metric is quite conservative as when computing the intersection, even single misaligned voxels can significantly reduce the metric value. If we, in fact, loosen the criterion of the objective function and we require only 90% of the corresponding voxels to be overlapping (as opposed to all), the overlap score increases



(a) Before congealing (b) After congealing

Figure 5-16: Synthetic data set of 40 MRI volumes: orthogonal slices of the mean volume of the samples (a) before and (b) after the stochastic congealing process.

to 98.52%. Second, interpolation artifacts also contribute to the reduction of our overlap score. As we compare binary segmented images that were transformed twice through the experiments, the partial voluming effect (partial assignment of integer valued labels to a voxel) can also result in slight mis-matches in the final result.

In order to convince the reader of the quality of alignment we also provide an image of the aligned overlapping ICC maps. Figure 5-17 displays the slices of the overlapping ICC volumes after alignment. According to the color-map, deep red corresponds to full alignment and blue to background. It is clearly visible that it is only on the boundaries, in a very narrow band, that there is any kind of deviation from the perfect alignment after registration. These errors are indicated by a yellowish color.

We also present the figure of the sum of entropies registration objective function computed along the optimization process. The two curves correspond to the improvement of the alignment scores obtained at two consecutive processing levels. Figure 5-18 (a) corresponds to the alignment metric computed on the lower resolution and (b) on the higher resolution level.

Error Analysis

Besides computing the overlap measures, we also analyze the transformations resulting from the congealing process. We are interested in knowing how closely our algorithm recovered the inverses of the transformations that created the synthetic data set.

In the case of congealing, computing such error measurements cannot be done directly. That is because we can only expect to recover the solution up to a com-

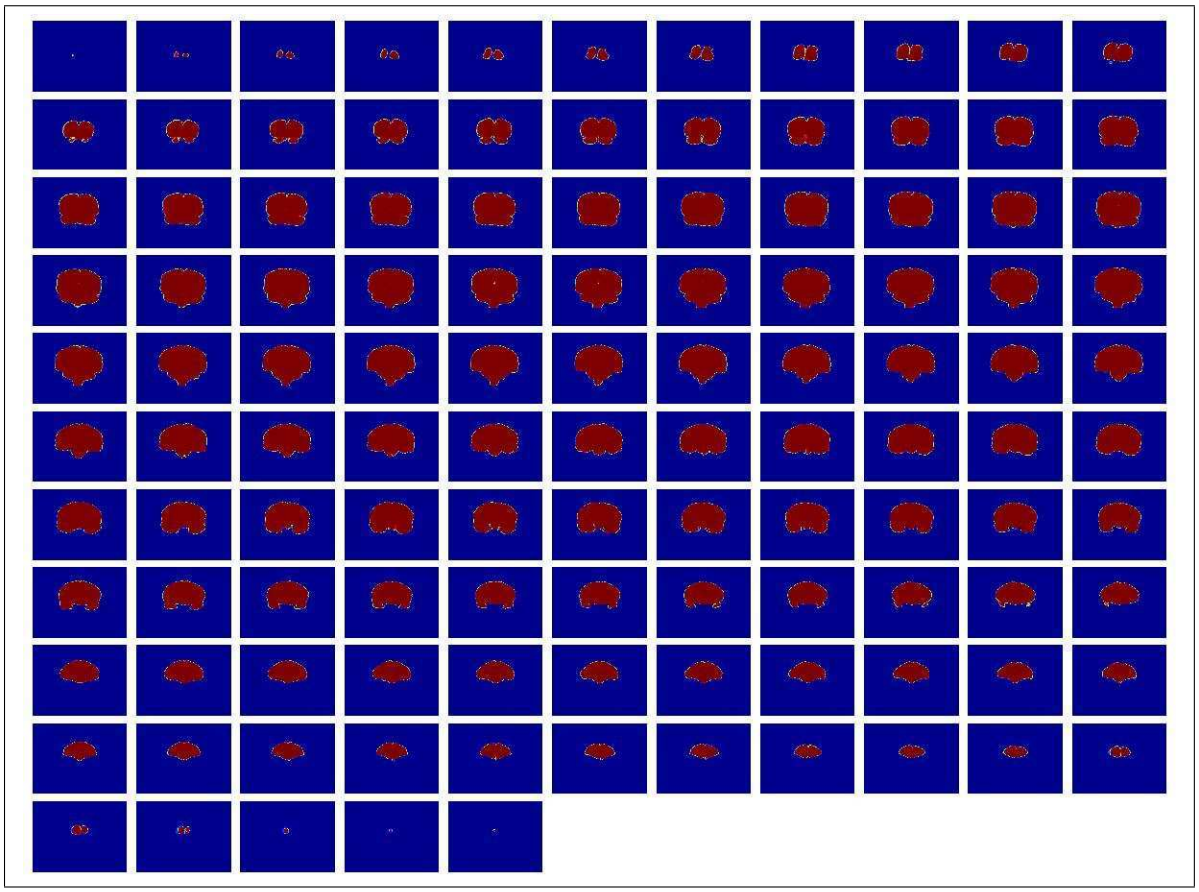
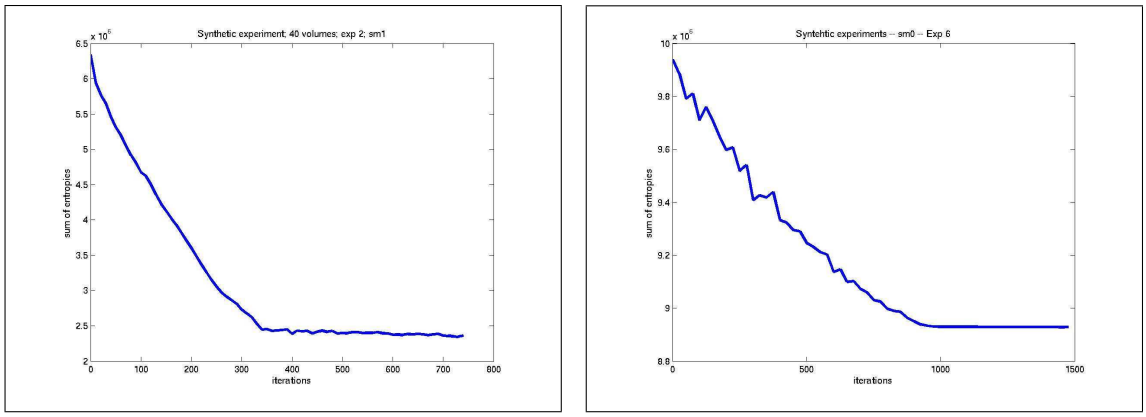


Figure 5-17: The slices of the ICC mean volume that was created after the results of the synthetic stochastic congealing algorithm were applied to the original segmentations. In the figure, deep red indicates 1 and deep blue indicates 0. The color of the ICC volumes is uniformly deep red, indicating success in registration. Any kind of slight disagreement (not perfect overlap) is indicated by a yellowish color. This only occurs on the boundaries and its size is so small that it is hardly visible. Such a discrepancy most probably results from interpolation artifacts.



(a) On Hierarchy Level 2 (low res.) (b) On Hierarchy level 1 (high res.)

Figure 5-18: The evolution of the congealing objective function throughout the optimization procedure of the synthetic data set consisting of 40 misaligned adult volumes. The sum of entropy plots were obtained on (a) hierarchy level 2 (the lower resolution data sets) and on (b) hierarchy level 1 (the highest resolution data sets).

mon transformation term. As we have hinted in Sec.4.2.6, given an arbitrary transformation component composed with each of the transformations recovered by the registration algorithm, the same registration results are produced as with out the composition. Therefore, simply reporting the difference between the inverse of the (known) offsetting ground truth and the recovered transformations would not provide a good characterization of our results.

Instead, we decompose the recovered transformations into a dispersion and a bias term. Our analysis is similar to the consistency measures introduced in [30]. The former error component indicates the variance in the accuracy of the recovered transformations across all the inputs and the latter conveys information about the magnitude of the common term. Figure 5-19 displays graphically how to interpret these error terms. In our accuracy analysis, low dispersion results are desirable as that indicates the reliability of our estimated transformations. The magnitude of the bias terms is of less concern to us. If all the resulting transformations embrace a common term (in addition to the desired one), the registration quality would still be the same.

In the following, we briefly describe how we computed these error measures. For notation, we refer to Fig. 5-14, where we indicate the set of offsetting transformations that produced the synthetic data set by $\mathcal{O} = \{O_1, O_2, ..., O_k\}$ and the transformations recovered by the congealing algorithm by $\mathcal{T} = \{T_1, T_2, ...T_k\}$.

When computing the error metrics in the spatial coordinate space of our input data volumes, we compare the transformation composition $C_i = (T_i \circ O_i)$ to the identity transformation T_I. Ideally, that composition produces the identity transformation, as the registration algorithm aims to recover the inverse of the set of offsetting transformations. In the case of congealing, because of the common component, we

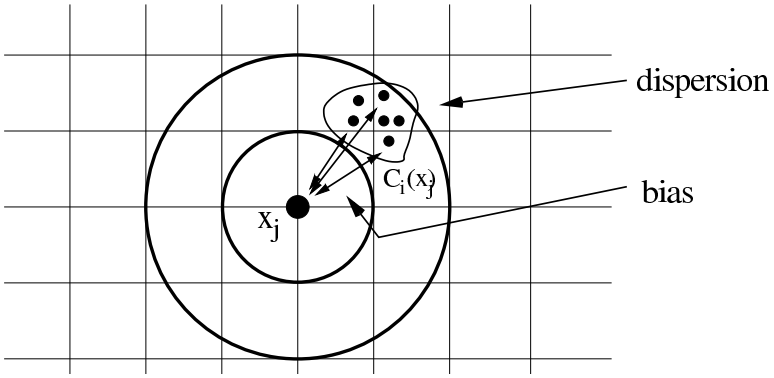


Figure 5-19: Graphical display showing how the dispersion and bias metrics are defined. The former describes overall variance in the transformed data locations and the latter describes the average magnitude of the difference from the true solution. We are interested in the former error component when validating our experimental results.

might not achieve identity even if the alignment is perfect. Thus, instead of directly comparing the identity transformation with the C_i transformations, we compute a bias and dispersion error term (or accuracy and precision, respectively). More specifically, we obtain these measures by first applying the composition transformation to all spatial coordinate locations. Given x_j, a spatial coordinate, the transformed point is at $x'_j = C(x_j)$. The bias term is concerned about the average magnitude of the distance between this and the original locations, and the dispersion term defines the variance of the location of the x' coordinates. Thus is we let $L_{ij} = \|x_j - C_i(x_j)\|$ indicate the length of the error vector, we can define:

$$\mathrm{Error_{bias}(x_j) = E[L_j] \quad and \quad Error_{dispersion}(x_j) = Var(L_j)}$$

Figure 5-20 displays our results. Given the set of forty transformation recovered by our congealing experiment, we computed a map of bias and dispersion terms. According to the corresponding color maps, both of the maps indicate low values. We are especially pleased as it is the dispersion terms that are the lower ones in the range of $[0, .3]$. In the center of the display, where most of the head data appeared, the values are the lowest. The bias terms (indicating the magnitude of the common term) were not too overwhelming either. They are between $[0, 2]$ voxels in most of the analyzed field.

The circular shape of the error maps in Fig. 5-20 can be explained by the fact that rotation is applied around the center of the input data set. Also, we associate the slantedness of the bias map with shear and scaling components.

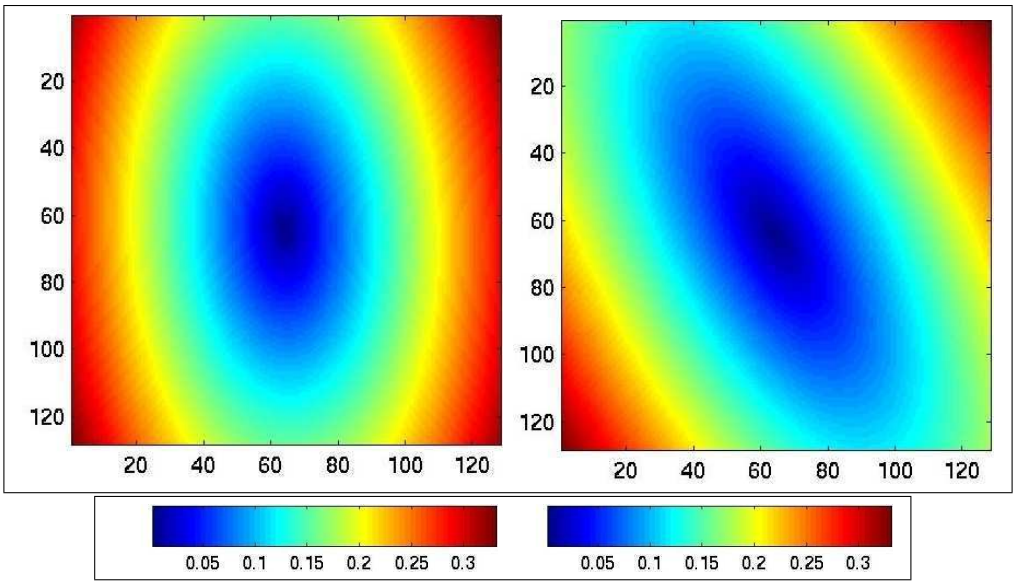


Figure 5-20: The (a) dispersion and the (b) bias maps of the synthetic experiments built in the spatial coordinate domain of the input data volume.

Repeatability Study

We recovered similar dispersion and bias fields when studying the repeatability of our group-wise registration results. We selected a data set of 25 T1-weighted MRI scans and repeated their group-wise registration 45 times. Before each experiment though, following the initial one, we randomly offset the input volumes. We then measured how similar the original and the post-offset registration results were. Figure 5-21 displays the outcome of one of our experiments. The dispersion measure is in the range of $[0, 2.5]$ and the bias measure in the range of $[0, 1]$. The former values are higher than in the case of the accuracy experiments and the latter are smaller. We believe that the increased range for the dispersion results can be explained by the fact that in this setting the congealing results are not compared to a ground truth; instead they are compared to one another. With respect to slightly different initialization settings, the alignment of the individual data sets also varies. However, that does not results in decreased accuracy (as shown by the accuracy experiments).

5.3.2 Mean Volume Atlas Comparison

In this section we briefly define how a mean volume representation of the stochastic congealing results compares to another method currently in use at our collaborating hospital. We compare the overlap measure corresponding to set of adult MRI head scans aligned by our method and by that of our colleagues'.

Warfield et al. applied a preliminary version of the original congealing approach to the problem of fusing MRI scans [79]. In their implementation, it was the pre-segmented intra-cranial cavity (ICC) of input volumes that was used for binary group-

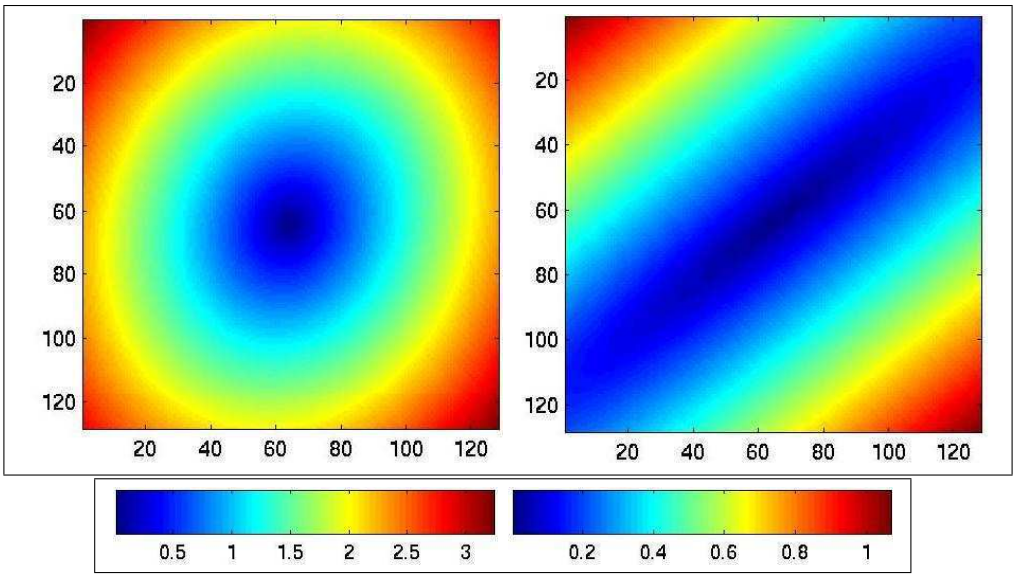


Figure 5-21: The (a) dispersion and the (b) bias maps resulting from the repeatability experiments.

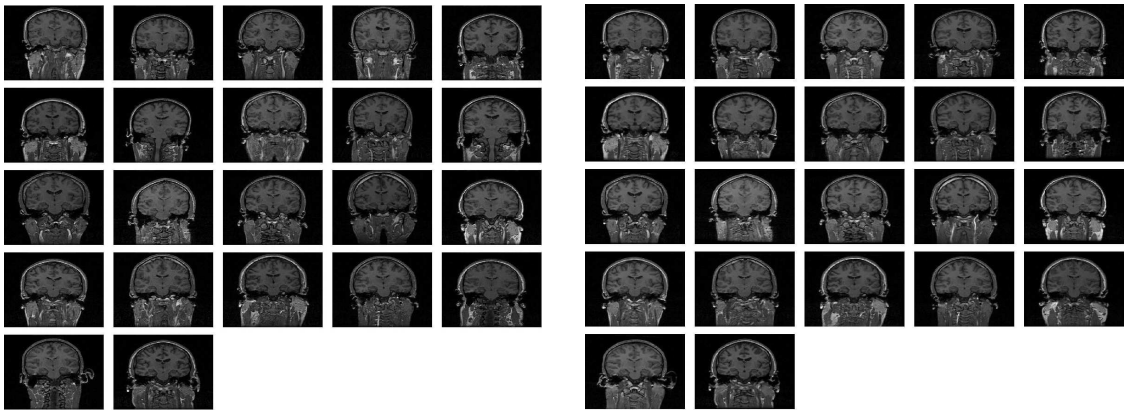
wise registration. One of the members of the group was also defined to be fixed. Then a nine parameter affine transformation (excluding shear) was identified for all but one of the inputs.

We present the the mean volume representation of the results of this method on adult brain scans in Fig. 5-23 (b). The volume itself is below referred to as the *control model*.

We compared the quality for the *control model* and for the mean volume obtained by our group-wise registration framework. More specifically, we ran our algorithm on the same set of twenty-two adult brain volumes on which the control model was defined and compared their overlap metrics.

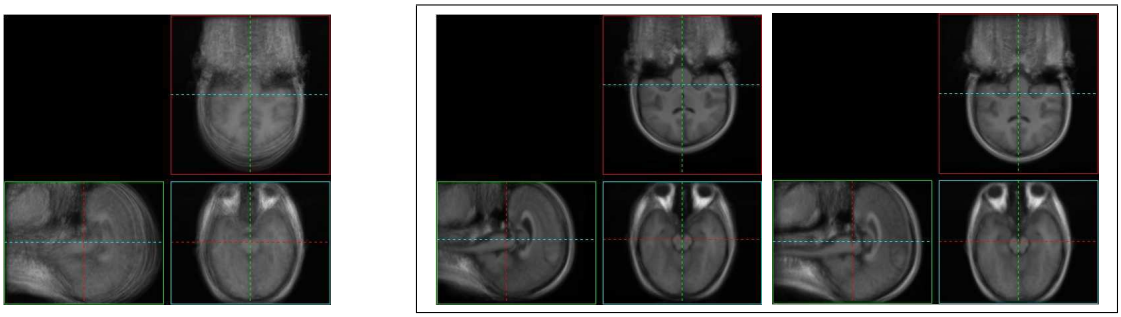
We first assess the success of the congealing algorithm qualitatively. Figure 5-22 displays the central coronal slices of the twenty-two input data volumes both before and after the alignment. Then, we can compare the quality of the mean volumes, or the atlases created by our and the control algorithm in Fig. 5-23 (b) and (c). We can establish that the clearness and the quality of these means are highly similar. For better appreciation of the results, the 3D view of the mean volume of the unaligned data set is demonstrated in Fig. 5-23 (a).

For a quantitative analysis, we use the same segmentation-overlap study as in the case of the synthetic experiments. We compute the overlap metric generalized from Eq.(5.1) for the ICC volumes of the two models, ours and the *control* one. With our registration results we obtain 86.33%, while with the control algorithm we achieved 84.48% overlap. Although the overlap metrics are very close, our alignment did better according to this metric.



(a) Before congealing (b) After congealing

Figure 5-22: The adult brain data set of twenty-two MRI volumes that were used to make our atlas. Central coronal slices of the input images (a) before and (b) after the stochastic congealing process.



(a) Before alignment (b) Control model (c) Our *atlas*

Figure 5-23: Three orthogonal views of the mean volume created from the adult brain data population of 22 images: (a) before alignment (b) the *control* model and (c) the mean model estimated by the stochastic congealing process.

The reader might notice that the aforementioned overlap measures are lower than in Section 5.3.1. That is because in the current experiment it is inter-subject scans and not synthetic ones that have to be aligned. The normal variability within the inter-subject scans can only be explained to a certain extent by affine transformations. The rest of the differences could be recovered by free-form deformations.

5.4 Pair-wise vs. Group-wise Image Alignment

When we described the main motivation behind studying group-wise registration algorithms, we argued that in order to align multiple number of data sets it is more advantageous to treat them as a collection as opposed to individuals. That is to say, we benefit from doing group-wise alignment as opposed to repeated pair-wise registration methods as the former is less susceptible to getting trapped in local optima and is more robust with respect to noise, truncation and occlusion artifacts. Similar

conclusions have been reached by other groups [74].

To justify this argument, we designed the following set of experiments. Given the adult brain MRI data that was used in Sec.5.3.2, we executed pair-wise alignment processes between all existing pairs of the input. Thus for a data set of $n = 22$ inputs, we ran $\left(\frac{[n*(n-1)]}{2} = 231\right)$ experiments. The pair-wise registration algorithm that we used for these 3D-to-3D experiments is mutual information-based [85] and it recovers 12-parameter affine transformations. The search space is explored via a gradient ascent optimization framework.

We found that in approximately half of the pair-wise registration cases the algorithm had to be rerun and the parameters readjusted. That was the case especially with the more challenging examples, when one of the images contained significant artifacts. Such a parameter readjustment was not necessary when running group-wise registration experiments as the framework is more robust towards outliers. Thus we may conclude from these experiments, that when the registration of a large number of input data sets is required, it is more advantageous to apply group-wise registration methods rather than executing pair-wise experiments.

5.5 Using Free-form Deformations

Although the 12-parameter affine transformations prove to be sufficient for many applications, they cannot explain all the differences that exist among inter-subject data sets. In order to eliminate some of the remaining local disagreements in our alignment results and to improve the quality of our registration, we implemented a free-form deformation framework. Such a deformation step can be introduced on the top-most level of our image processing pyramid where it is the highest resolution form of the data sets that is processed. As opposed to just recovering twelve transformation parameters, the computation of a dense deformation field requires the optimization of thousands of parameters per input.

We implemented approximations to two widely used deformation formulations: the *viscous fluid* and the *linear elasticity* models. The former allows for large scale deformations, while the latter is more restrictive and is more appropriate for recovering small scale warps. We implemented a kernel-based fast estimation of these methods [4, 15]. Briefly, the computationally expensive, sequential over-relaxation step in the original framework of Christensen [9] is replaced by a filtering step. More specifically, the application of a Gaussian filter on the gradient-based update field results in a viscous fluid-type warp. On the other hand, if the filtering step is applied to the update terms and not to the deformation field, then the warp resembles the linear elasticity model. Computing the non-rigid warps according to these two frameworks is summarized in pseudo-code in Alg. (2) and Alg. (3).

Note, that in the case of the viscous fluid model there is an explicit regularization step that is required. It computes the sign of the determinant of the Jacobian

Algorithm 2 Top-level code describing one iteration of the free-form deformation computations according to the viscous fluid model.

```
for all input data volumes do
  Compute gradient-based update terms to deformation field
  Add all the updates to the deformation field
  Apply smoothing kernel to the updated deformation field
  if Negative determinant of Jacobians then
    Re-grid and reset the corresponding deformation field
  end if
end for
```

throughout the whole deformation field. If at any location that value turns negative, the deformation field is re-gridded, following the description in [9, 4]. Under the linear elasticity model, the smoothing operations are sufficient to serve as regularization terms [62]. Further implementation details and experimental results are described in Chapter 6.

Algorithm 3 Top-level code describing one iteration of the free-form deformation computations according to the linear elastic fluid model.

```
for all input data volumes do
  Compute gradient-based update terms to deformation field
  Apply smoothing kernel to the set of updates
  Add all the updates to the deformation field
end for
```

As the computation of the free-form deformation parameters significantly increased the total computation time of our experiments (approximately 2 days for a data set of 22 volumes), we present preliminary 2D warping results. As the quality of the linear elastic and the viscous fluid were comparable, in Fig. 5-24, we only show results from the latter type of experiments. More specifically, in the figure, we display three members of the data set before and after the warping experiment. The red contour indicates the final outline of the mean image of the corresponding slices in order to better appreciate the magnitude of the individual deformations. We may establish that all three of the examples converge towards each other and the outline of the mean slice is more in agreement with the individual slice boundaries after the warping process. Note, that the experiments, in the future, will need to be executed also on a higher resolution grid. Even though the current warp allowed the shape of the brains to be properly overlapping, it was not completely sufficient to achieve gyral correspondence.

An additional difficulty in the case of non-rigid registration experiments originates from the fact the the input data sets have to be skull-stripped. That is to say the skull and CSF layers need to be identified and deleted from the images in order to allow for an emphasis of the precise alignment of the cortical structures [60]. Even though

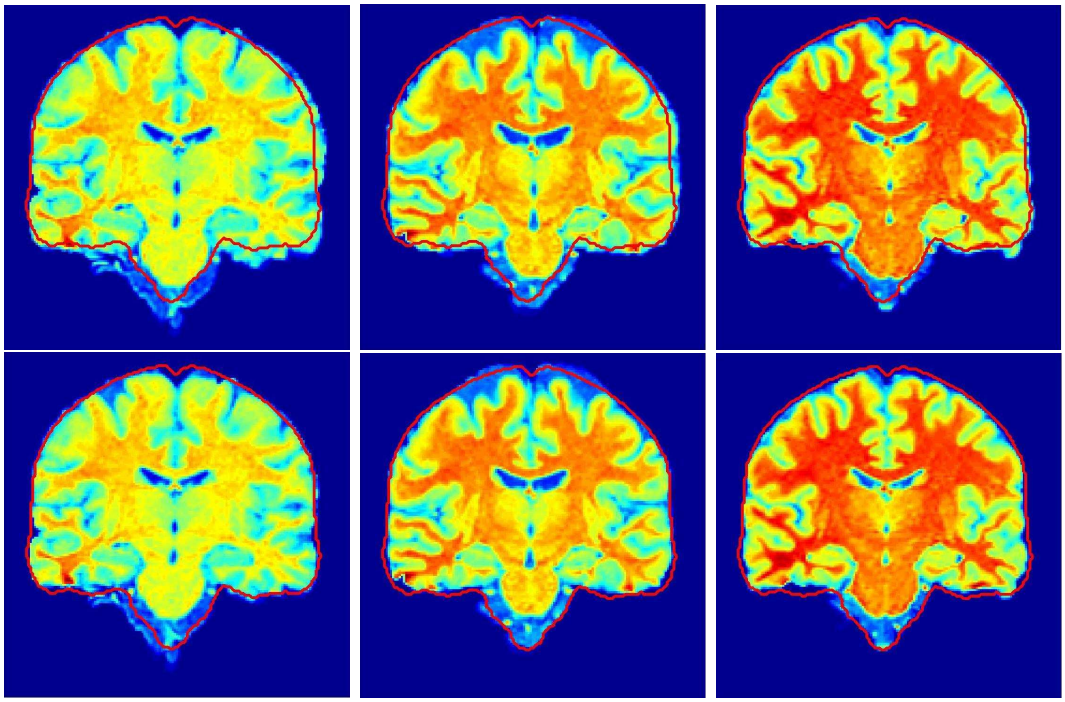


Figure 5-24: Warping results from 2D experiments. Three members of the data set (top) before and (bottom) after the warping experiment. The red outline indicates the final outline of the mean image of the corresponding slices.

this preprocessing step is often needed in many medical imaging applications, it is challenging to find a robust implementation that does not require significant human supervision. Any segmentation artifacts that remain as a result of such a procedure might negatively affect the performance of the subsequent registration procedure.

5.6 Summary and Conclusion

In this chapter we demonstrated the performance characteristics of the stochastic congealing algorithm through both qualitative and quantitative analysis of experiments. We provided results on a wide variety of data sets, and we also derived quantitative error measurements to describe accuracy and repeatability features. The excellent results from the large population affine experiments allow us to for reinforce the attractive properties of our novel group-wise registration framework. In the upcoming chapter, we present a two additional experimental results. These show how our registration results can be combined with hypothesis testing and segmentation procedures in order further analyze the collection of input data sets.

Chapter 6

Additional Applications

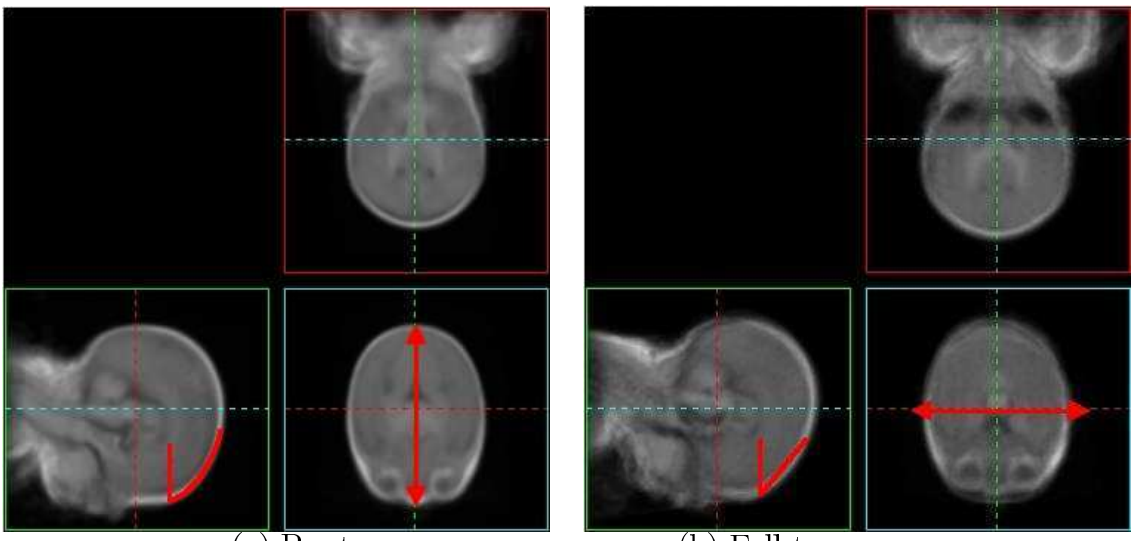
In this chapter, we demonstrate two additional applications that greatly benefit from the results of the stochastic congealing group-wise registration method. We first introduce a group analysis framework which allows us to recover sub-population characteristics from the set of congealing transformations and then we demonstrate how one might use stochastic congealing in order to improve the segmentation quality of baby MRI scans by creating unbiased label probability maps.

6.1 Characterizing Sub-Populations by Joint Alignment

Examining a set of pre- and full-term baby brain MRI data, our medical collaborators established that the two populations are likely to demonstrate systematic differences in the shape of their skulls [46]. The most visible changes appear in the amount of elongation visible in the axial scans (in the case of the pre-term population the head is more elongated) and also in the curvature of the forehead (higher curvature surface for pre-terms and more triangular shape in the case of the full-terms). In Figure 6-1 we show these proposed differences on the mean volumes of the separately aligned populations.

In order to establish whether these differences are statistically significant though, we have to analyze the data sets in the same coordinate system. That not only prompted us to use our stochastic congealing method for the group-wise alignment of those volumes, but also to address a more general problem. We were interested in investigating whether it was possible to characterize sub-populations within a large data set, or in other words, whether it was possible to identify multiple central tendencies in the original input data set by studying the distribution of transformations resulting from the population alignment.

The pre-term data set that we use in our experiments has been already described

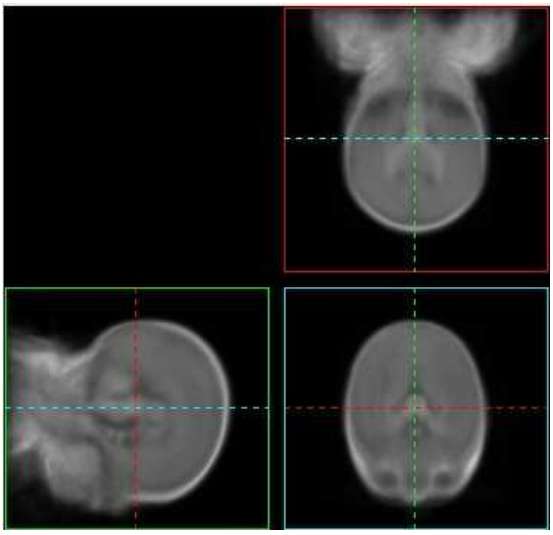


(a) Pre-terms (b) Full-terms

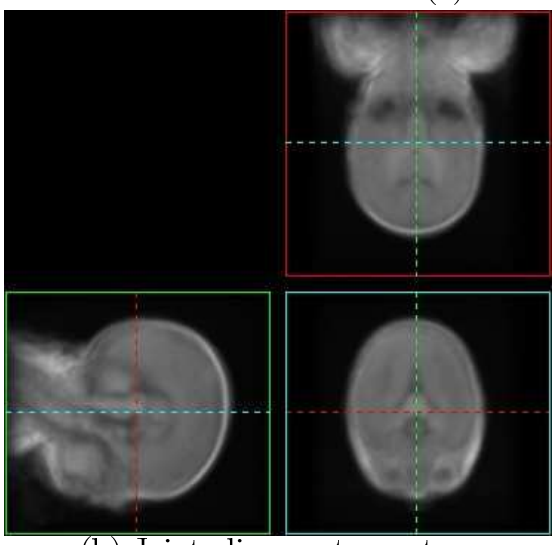
Figure 6-1: Two of the key differences in the head-shape of the pre-term and full-term baby brains that are thought to be characteristic: the curvature of the forehead and the elongation of the skull in the axial view.

in Sec. 5.2.1 and we characterize the full-term data collection in the following. The full-term baby data set consists of the acquisitions of seventeen different subjects. The volumes are 256 by 256 by 110 voxels, with each voxel measuring 0.703125 by 0.703125 by 1.5 millimeters. Although T1-, T2- and PD-weighted MRI scans are available for all, in the following experiments we only refer to the T1-weighted images.

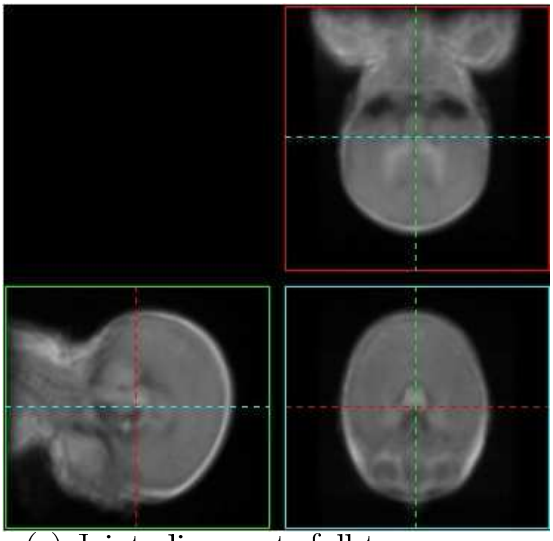
First, we jointly aligned the two groups of pre- and full-term data sets. The thus created mean volume does not seem to favor either of the distinguishing shape characteristics as it is displayed on Fig. 6-2 (a). Given that the sub-population sizes are comparable, we expected that result. Then, as a preliminary experiment, we again created the mean volumes associated to the two sub-groups. This time, however, we used the transformations recovered by the "joint" group alignment within each. After registering the pre- and full-term populations as one group, we created two mean volumes by separating the aligned input data sets according to their labels. Although the distinguishing features between the two sub-populations were not as pronounced as before, the differences were still visible (see Fig. 6-2 (b) and (c)). These results made us further believe that the mentioned sub-group differences can indeed be recovered.



(a) Joint alignment



(b) Joint alignment: pre-terms



(c) Joint alignment: full-terms

Figure 6-2: Three orthogonal views of the mean of congealed volumes. The joint data set of pre- and full-term scans was congealed together as one set and then the group means of the joint, the pre- and the full-term volumes were constructed. Figure (a) demonstrates the joint, (b) the pre-term and (c) the full-term mean.

6.1.1 Permutation Testing

In order to validate the significance of our observations with respect to the baby MRI population, we formulated the problem in a *statistical hypothesis testing* framework. This provides a principled approach to test alternatives for and against assertions that were made based upon observations of a set of samples. When using hypothesis testing, one needs to define a *null* (H0) and an *alternate* (H1) hypothesis about the observations. Briefly, the former claims that any kind of observable differences are the result of pure chance and according to the latter the observations show a real effect combined with a component of chance variation. The goal of the hypothesis testing is to gather enough evidence to reject the null hypothesis in a statistically significant manner [16].

Our null hypothesis proposes that there is a single underlying distribution that characterizes the input data set and thus the combined population cannot be separated. The alternate hypothesis claims that the two populations have different distributions. In order to test for the null hypotheses, a test statistic needs to be identified. The distribution of such a metric is then evaluated under the null hypothesis. The critical value of the test allows us to decide whether the null hypothesis can be rejected. In our case though, we did not have access to the required distribution under the null hypothesis. Thus we used the *permutation testing* method [22].

Permutation testing is a non-parametric technique for hypothesis testing. It estimates the probability distribution of the statistic under the null hypothesis from the available data. Random mathematical permutations are applied to the labels of the observations. After re-assigning these labels the test statistics are re-computed. That process is repeated a large number of times. Finally, the test statistic for the true labelling is compared to all the other values in deciding about its significance.

Permutation tests are special cases of randomization tests, i.e. tests that use randomly generated numbers for statistical inference. They operate under the assumption that the data distribution is adequately represented by the sample data. In our experiments, we have not addressed the question of whether the sample data adequately describes the population.

According to the sampled or approximate permutation test, a test proceeds as follows:

I. Combine the observations from all the samples.

II. Permute the labels of the observations and redistribute them among the original data samples.

III. Compute the statistic of interest.

IV. Repeat steps (2) and (3) sufficient number of times.

V. Determine how often the re-sampled statistic of interest is as extreme as the observed value of the same statistic.

This procedure computes an empirical estimate of the cumulative distribution of a statistic under the null hypothesis and uses it for hypothesis testing. Since the null hypothesis assumes that the two classes are indistinguishable with respect to the selected statistic, all the training data sets generated through permutations are equally likely to be observed under the null hypothesis, yielding the estimates of the statistic for the empirical distribution.

Ideally, we would like to run an exhaustive permutation test, that is to say we would prefer to incorporate the results of all possible shufflings in the decision making. However, that might not always be feasible because of computational limitations. In our experiments, we randomly sample from the pool of all possible permutations. Thus it becomes very important to select the number of sampling iterations to be large enough to guarantee accurate estimation. In our experiments we used N = 12000 random permutations for obtaining our results.

6.1.2 Test Statistic

A test statistic must be chosen in such a way that it is relevant to the hypothesis and it summarizes the important information in the observed data samples.

As opposed to defining test statistics in the domain of the observed and aligned images, we decided to construct our test metrics based upon the set of transformations that have been recovered by our group-wise alignment process. In the case of the baby MRI data set, we expected to see significant scaling and shearing differences in the aligning transformations if the observed skull shape differences were real. As the displacement and the rotation transformations are not informative, we discarded them. We then defined a set of three *scaling ratios*[1] extracted from the mean transformation of each subgroup. Intuitively, we expected that one or more of these metrics (especially the ones containing the axial direction information) is going to be able to capture the observed differences among the input data samples.

In order to compute the average transformation corresponding to a subgroup of the input data sets, we used the square Frobenius norm in order to establish *distance* between two transformations. That norm is defined as the square root of the sum of the absolute squares of its elements, or if A is an (m x n) matrix, then

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}. \tag{6.1}$$

[1]We call these metrics *scaling metrics* for simplicity, but it is possible that they also contain shearing components. The decomposition of an affine transformation matrix into scaling, rotation and shearing components, in general, cannot be done uniquely.

As a matrix distance, we used $D_F(A,B) = \|AB^{-1} - I\|_F$ where A and B are matrices of the same dimensions and I is the identity matrix. We defined the mean of a group of k transformations ($\mathcal{T} = [T_1, ..., T_k]$) according to:

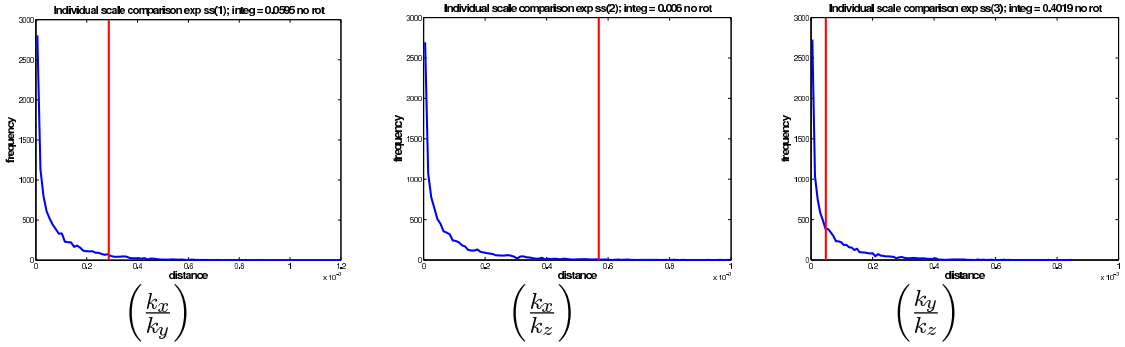
$$\hat{T} \equiv \arg\min_{T} \sum_{i=1}^{k} D(T, T_i). \tag{6.2}$$

If we set $x_0 = [0,0,0]$ and $(\hat{x}, \hat{y}, \hat{z}) = \{[1,0,0]; [0,1,0]; [0,0,1]\}$, our three scaling ratios are summarized in $\mathcal{S}$:

$$k_x = |\hat{T}(\hat{x}) - \hat{T}(x_0)|; \quad k_y = |\hat{T}(\hat{y}) - \hat{T}(y_0)|; \quad k_z = |\hat{T}(\hat{z}) - \hat{T}(z_0)|; \tag{6.3}$$

$$\mathcal{S} = \left(\frac{k_x}{k_y}, \frac{k_x}{k_z}, \frac{k_y}{k_z}\right). \tag{6.4}$$

6.1.3 Experiments

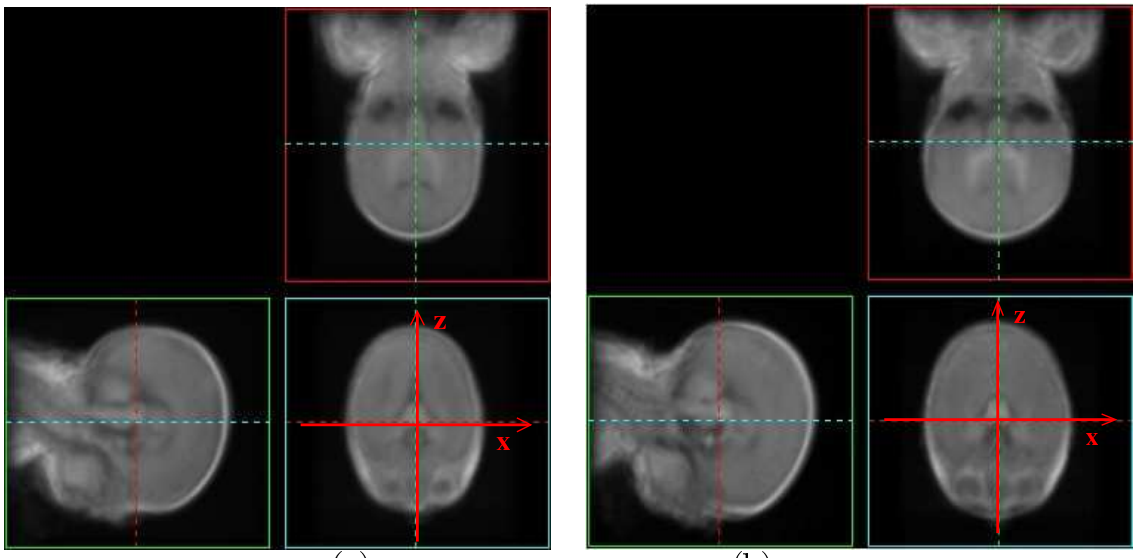


$\left(\frac{k_x}{k_y}\right)$ $\left(\frac{k_x}{k_z}\right)$ $\left(\frac{k_y}{k_z}\right)$

Figure 6-3: The test statistic distributions for (a) $\left(\frac{k_x}{k_y}\right)$ (b) $\left(\frac{k_x}{k_z}\right)$ and (c) $\left(\frac{k_y}{k_z}\right)$ attained after running our permutation testing. The red vertical lines in the graphs indicate where the values of the metric would lie when computed with the true labelling.

After running the congealing algorithm on the joint set of pre- and full-term baby MRI data volumes (where the total number of volumes was 37), we recorded the resulting transformations and ran the permutation test analysis for $N = 12000$ iterations. Figure 6-3 displays the distribution of the test statistics generated by the permutation test analysis. For the three scaling components respectively, we obtained a critical value of (.0595, .006, .4019). These are the values of the metric when it is computed with the true labelling. On Fig. 6-3, it is the red vertical lines that indicate these values. At the level of $\alpha = .05$, we consider the second ratio comparing the x and z directions to be significant. As the critical value of .006 is considerably smaller than α, we may establish that there is strong evidence that the null hypothesis is wrong and it can be rejected.

Figure 6-4 demonstrates the dimensions along which we were able to separate the subgroups.



(a) (b)

Figure 6-4: Three orthogonal views of the mean of congealed volumes of the (a) preterm and the (b) full-term data sets using transformations recovered by the combined group-wise alignment of these two sub-populations. The red axes indicate the two directions whose ratio provided the most significant statistic in characterizing the two sub-populations.

This initial finding was very satisfying. According to an on-going collaboration with the Brigham and Women's Hospital, these distinctions between the sub-populations is medically expected and can be explained. After they are born, preterm babies lie in incubators as opposed to "floating" in the womb. Lying then on the side of the head, can significantly affect the head shape [46].

6.2 Segmentation with Unbiased Atlases

The other application that we present in this chapter is related to medical image segmentation. The task of this image processing algorithm, as we have briefly explained it in Chapter 1, is to assign descriptive labels to spatial locations in the data coordinate system. These labels frequently correspond to anatomical features or tissue types. Segmentation is often a very challenging task as the image intensity values themselves might not be directly converted into classification labels. In general, intensity values describing a particular tissue type can vary in a wide range, they might be overlapping the intensity range generally associated with another label and they may also be distorted due to imaging and motion artifacts. Thus additional information about the input or the imaged anatomy is often used in order to guide the segmentation process. Manual segmentation (classification done by human experts), for example, relies on the knowledge of anatomy of the segmenter. Although it is

often considered to be a gold standard, manual segmentation is very labor- and time-intensive. Hence, for group studies, automated or semi-automated procedures are more preferable.

In Chapter 1, we also mentioned that segmentation and registration are two closely coupled medical image processing tasks. Their roles are complimentary, given that the result of one can significantly improve that of the other. Frequently, these two processes are either executed simultaneously or run sequentially. In this section, we demonstrate how we used the stochastic congealing group-wise registration results to improve the performance of a segmentation algorithm on a set of challenging MRI data sets.

6.2.1 The Segmentation Problem

A particularly difficult segmentation problem is the labelling of neonatal data sets. Our collaborators wanted to identify cortical gray matter (cGM), myelinated and unmyelinated white matter (mWM and uWM), cerebrospinal fluid (CSF), basal ganglia structures (BG), and extracerebral tissue based upon MRI acquisitions. That is a particularly complex task due to reduced MRI signal-intensity contrast between tissues of interest compared to older subjects. Although each data set had already been automatically segmented at image acquisition time, an improved classification was needed in the case of almost all the data sets as the initial ones were generally of low quality. One example of an inaccurate segmentation, for example, is displayed in Fig. 6-5. The figure displays an axial slice from one of the head scans where the colors correspond to different tissue labels. Here, both of the eyes are mistakenly identified as white and gray matter.

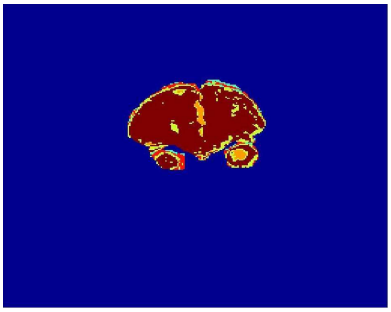
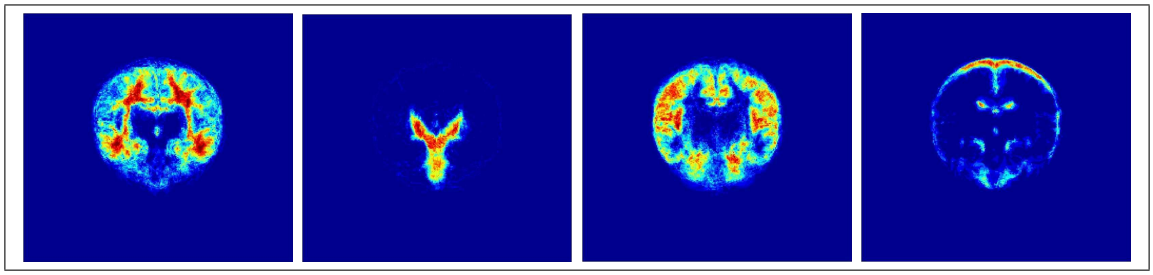


Figure 6-5: An axial slice of a classification map: the eyes are identified as gray and white matter.

One way to introduce additional information into the segmentation procedure, besides requesting manual labelings, is to use label probability maps. Such maps are constructed from a collection of previously observed segmentation examples. After the individual cases are positioned in the same coordinate system, we register the frequency of a label occurring at a particular voxel location given all the observations. In Fig. 6-6, we demonstrate the central slices of four such label maps corresponding

to the pre-term baby data set: uWM, mWM, cGM and CSF. The color values in these images cover the $[0, 1]$ range. Dark blue and dark red are assigned to the lowest and the highest extremes. The higher these values are, the higher percentage of the aligned data sets agreed on the classification of that data point.



(a) uWM (b) mWM (c) cGM (d) CSF

Figure 6-6: Label probability maps computed following the group-wise alignment of the pre-term data set. The indicated tissue labels correspond to: uWM - unmyelinated white matter; mWM - myelinated white matter; cGM - cortical gray matter; CSF - cerebro-spinal fluid. The color code covers the [0,1] range with dark blue indicating 0 and dark red indicating 1.

The way these label maps are created might be significant when evaluating their influence on further segmentations. Such collection of prior information might guide segmentation methods in challenging situations, but it might introduce noise if not properly created. In this section, we propose to use the stochastic congealing method to produce an unbiased set of label probability maps and compare its power to other such representations.

6.2.2 The Training and Test Data Sets

We used T1-weighted MRI volumes of twenty healthy pre-term children in our *training set* to develop label probability maps corresponding to four labels. These were initially segmented by an expert reviewer using a previously published supervised classification system [81]. Based upon anatomical correspondences computed via our group-wise registration algorithm, the segmentation maps were then also brought into alignment. Figure 6-6 displays the one slice of each of these label maps.

The *test set* of the segmentation experiments comprised the MRI volumes of five new pre-term subjects (that were not part of the training set). Orthogonal slices of each member of this challenging data is displayed in Fig. 6-7. These scans were each associated with five corresponding segmentations labelled semi-automatically by different experts. Thus, overall, we had twenty-five segmentation configurations with which we could measure the segmentation performance.

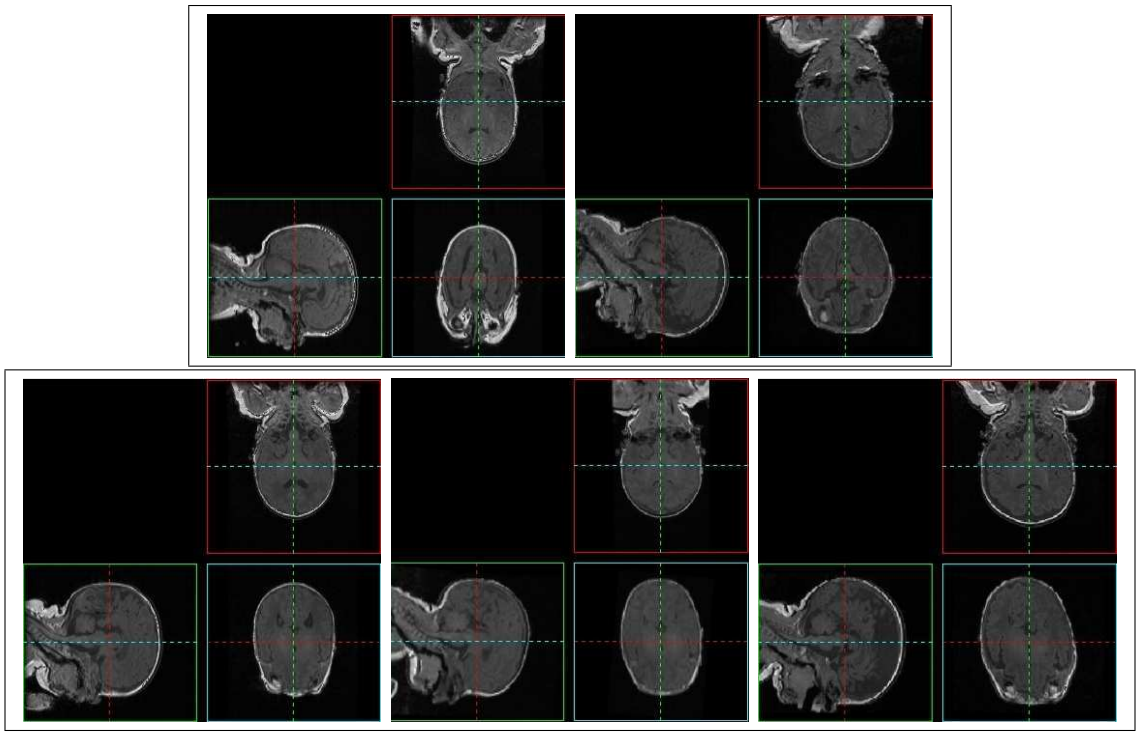


Figure 6-7: The five T1-weighted MRI scans whose segmentations we used for evaluating the use of atlases.

6.2.3 Algorithm

As mentioned above, the overall quality of the original segmentation results were frequently poor. In order to eliminate or reduce the size of these segmentation errors, we propose to use the label probability maps computed after congealing. In order to evaluate the utility of such prior information and to compare the segmentation results quantitatively, we propose two directions for our analysis. First we analyze the variability produced by the guided segmentations and then we assess their accuracy. In the former study, we first show that using an atlas does significantly improve the reproducibility of segmentations compared to those that do not use prior information. Then we compare the performance of using our unbiased atlas as opposed to a biased one. In the accuracy experiments, we validate the newly proposed segmentation results and a previously published method from our collaborators' group against a group of manually drawn segmentations. The semi-automatic segmentation algorithm that we rely on during these experiments is a fast *k-nearest-neighbor* (kNN) segmentation algorithm [78], which is able to use atlas information. User input is required in the form of providing a set of *seed points* with a known label assignment. In order to be able to replicate the segmentation experiments in our analyses, we saved the seed points of all the original segmentations in our input data set.

6.2.4 Segmentation Variability

As obtaining a gold standard segmentation is not always feasible, we first use a quantitative evaluation algorithm called STAPLE[2] [80] to assess the variability and label consistency of image segmentation results. Given a set of segmentations of the same input, this algorithm simultaneously computes a reference standard (or estimated true segmentation) and also the performance level of the individual input classifications. Performance is described by sensitivity and specificity parameters, but we primarily report *predictive values* (PV) of the different segmentation labels. We compute their mean, standard deviation and coefficient of variation over all the subjects and all the labels. The mean predictive value is defined as the probability of having the true segmentation label agree with the predicted one. For a particular label l, the true label t and segmentation decision d, it is equal to $\mathrm{PV} = \Pr(t = l | d = l)$. Having, for instance, this indicator increase when using our unbiased atlas, would be one indicator that our registration results may improve segmentation results. The coefficient of variation (%) of a set of values is calculated as: $CV = 100 * \left(\frac{\sigma}{\mu}\right)$. A decreased value of CV would mean lower intra-segmentation variability, which is another desirable feature for large-scale segmentation studies.

Segmentation Without and With an Atlas

In the first set of experiments, we demonstrate the advantages of using prior knowledge encoded in label probability maps for segmentation. We ran the segmentation algorithm with and without using these label probability maps and summarize the results in the first two columns of Table 6.1.

The segmentation procedure that does not use any prior information performed very poorly and the one using the atlas was significantly better having a mean predictive value of .5366 and .9425 respectively. We do not want to overestimate the value of the outcome of these results. Using any type of relevant information might increase the accuracy of the segmentation, so it is slightly unfair to compare no atlas vs. atlas procedures. It is more for the sake of completeness that we describe these results .

Segmentation Using Biased and Unbiased Atlases

As mentioned above, in the second set of experiments we compare two segmentation methods that both use prior information about the spatial likelihood of classification labels. While the segmentation algorithm is the same, the nature of the prior information is different. In the case of one, a biased atlas is used and in the case of the other an unbiased. The former is defined with respect to the segmentation of a single volume and the latter to the mean atlas that is produced by stochastic congealing. The results of these experiments are included in the last two columns of Table 6.1.

[2]The acronym STAPLE stands for *simultaneous truth and performance level estimation*.

	No atlas	Biased Atlas	Unbiased Atlas
cGM	.65	.84	.96
CSF	.61	.88	.92
mWM	NA	.8	.94
uWM	.35	.9	.95
μ_{PV}	.5366	.855	.9425
$\mathrm{CV} = 100 * \left(\frac{\sigma_{\mathrm{PV}}}{\mu_{\mathrm{PV}}}\right)$	66.49	5.43	2.94

Table 6.1: Results of the segmentation experiments. The table contains statistics about the predictive value (PV) computed by STAPLE. The first five rows contain mean PV scores corresponding to specific labels. The sixth row indicates the mean PV (μ_{PV}) value across all the tissue labels, the seventh row indicates the standard deviation of PV (σ_{PV}) over all the tissue labels and the 8th row displays the coefficient of variance computed from the mean and standard deviation. The label acronyms correspond to: cGM - cortical gray matter; CSF - cerebro-spinal fluid; mWM - myelinated white matter; uWM - unmyelinated white matter; sGM - subcortical gray matter. NA corresponds to values that are not available.

Segmentation experiments using the unbiased label probability maps produce significantly better results, in terms of group consistency, when compared to those that rely on a biased atlas. The mean predictive value for all tissues for the 25 experiments increased by 8.75%. The variability of the individual segmentations (indicated by CV) also decreased from 5.43 to 2.94. These results are significant. They underline the fact that using an unbiased way of constructing an atlas retains more relevant information about the group than using biased approaches.

6.2.5 Segmentation Accuracy

With our collaborators, we ran another set of segmentation experiments in order to assess accuracy information about the congealing-guided segmentation. When measuring segmentation accuracy, validation was performed by comparing the segmentations to ground truth estimates in each of 5 subjects. An estimate of ground truth was provided by having an expert rater manually assign tissue labels to each pixel corresponding to a single MRI slice in each of our test subjects. More details about these experiments are described in [83].

In Fig. 6-8, we display the 2D segmentation results that were compared in the above study. Qualitatively we may claim that it is the segmentation method that uses the results of our group-wise registration algorithm that best resembles the manual segmentation outcome. In the figure, colors represent the following labels: cerebrospinal fluid (blue), myelin (orange), cortical grey matter (grey), basal ganglia (white), and unmyelinated white matter (red).

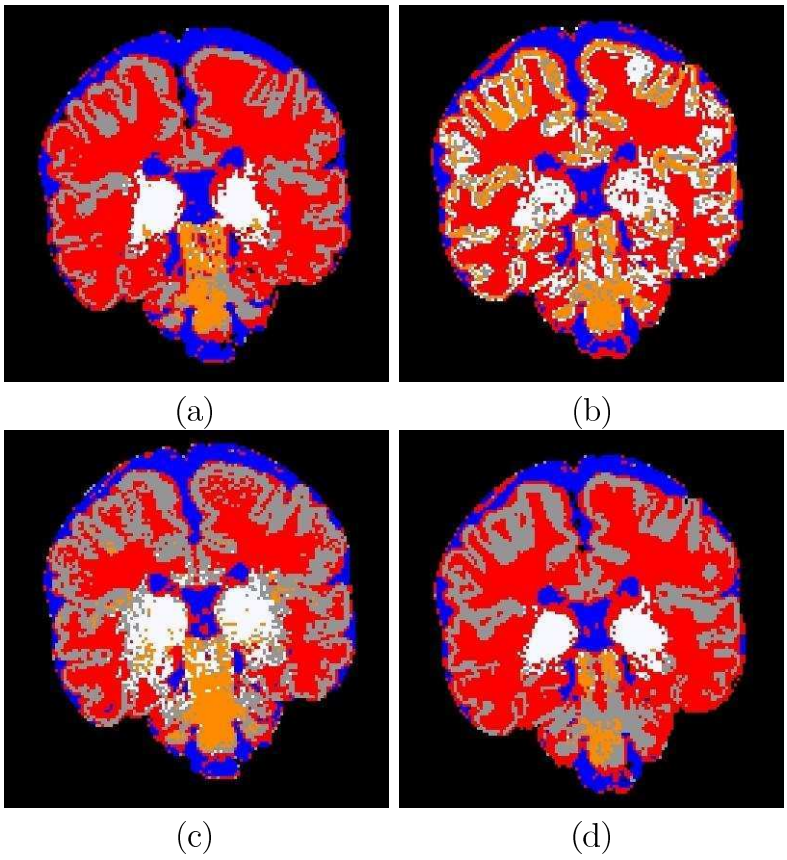


Figure 6-8: 2D segmentation results using: (a) expert-labeled manual segmentation (b) no atlas information (c) the old pipeline and (d) the congealed atlas. Colors represent cerebrospinal fluid (blue), myelin (orange), cortical grey matter (grey), basal ganglia (white), and unmyelinated white matter (red).

In order to quantitatively evaluate a segmentation result, we computed the Dice similarity coefficient ([91]) with respect to the manual segmentation. The previously reported biased segmentation (biased) [81], and the segmentation results using unbiased statistical atlases (unbiased) and no atlas at all (no atlas) are compared against each other. Our results are shown in Table 6.2 for each subject and each tissue class. Overall the method relying on the unbiased atlas equals or outperforms the previously validated method and achieves excellent or near-excellent results. Note that Zijdenbos claims that greater than 0.7 is considered excellent agreement [91] in the case of the Dice similarity metric. The no atlas results are again shown only for completeness.

Subject	Algorithm	cGM	CSF	mWM	uWM	BG
1	biased	0.63	0.76	0.67	0.68	0.68
	no atlas	0.23	0.77	0.48	0.76	0.35
	congealed atlas	0.76	0.66	0.67	0.77	0.74
2	biased	0.67	0.67	0.77	0.61	0.53
	no atlas	0.32	0.68	0.39	0.71	0.29
	congealed atlas	0.68	0.66	0.52	0.71	0.72
3	biased	0.65	0.69	0.76	0.70	0.69
	no atlas	0.30	0.72	0.33	0.73	0.35
	congealed atlas	0.77	0.71	0.77	0.77	0.73
4	biased	0.70	0.43	0.56	0.62	0.64
	no atlas	0.56	0.54	0.30	0.71	0.27
	congealed atlas	0.73	0.51	0.71	0.71	0.64
5	biased	0.71	0.71	0.67	0.65	0.72
	no atlas	0.48	0.69	0.39	0.72	0.32
	congealed atlas	0.81	0.70	0.75	0.72	0.76
mean	**biased**	**0.67**	**0.65**	**0.69**	**0.65**	**0.65**
stdev		0.03	0.13	0.09	0.04	0.07
mean	no atlas	0.38	0.69	0.39	0.72	0.32
stdev		0.14	0.09	0.08	0.02	0.04
mean	**congealed atlas**	**0.75**	**0.65**	**0.69**	**0.74**	**0.72**
stdev		0.05	0.08	0.1	0.03	0.05

Table 6.2: Comparison of three segmentation results with single-slice manual segmentation using Dice similarity coefficient. The three different segmentation methods are: (a) pipeline using biased statistical prior, (b) new pipeline using no statistical *prior* atlas; and (c) the new pipeline using the statistical atlas resulting from congealing.

In this section, we have shown the current work to equal or exceed the previous method in accuracy, on average, for five challenging segmentations. While the previous methodology took a trained reviewer several iterations to achieve adequate results, similar, and often better, results were achieved when using the unbiased atlas in far less time. The clinical significance of studies done with this technique, com-

bined with the current speed and ease of use, make it possible that this technique will find routine use in the clinical evaluation of premature infants at our collaborators' institution. Future work should include testing on a larger sample of children and further analysis of the effect of each stage of the processing pipeline.

6.3 Summary

In this chapter, we demonstrated two applications that benefit from the group-wise registration results of the stochastic congealing algorithm introduced as a pre-processing step. We described how sub-populations of larger data sets could potentially be characterized by examining the transformations recovered by the group-wise alignment technique. Additionally, we showed how our unbiased population registration results could be utilized in order to guide challenging segmentation tasks.

Chapter 7

Conclusion and Future Plans

7.1 Conclusion

In this dissertation, we studied the problem of pair-wise and group-wise registration of medical image data sets. We were interested in understanding the strength and weaknesses of currently available, registration approaches and also in defining new alignment approaches that improve on the existing results. Accordingly, we first construct a unified statistical framework in order to examine and compare a set of widely used pair-wise registration algorithms. Although the formulation of the framework originates from the classical maximum likelihood interpretation of image registration, we use a closely related information theoretic basis to draw our comparisons. That allows us to include a wider set of principled similarity measures into our analysis. In certain scenarios it is desirable to rely on prior information about statistical features of previously aligned data sets. While such knowledge provides advantages, it can also limit the accuracy of the alignment. Thus, instead of using a fixed model, we demonstrate how one can incorporate *a priori* knowledge into a registration framework by using a distributional assumption.

Besides the pair-wise registration problem, we also investigate group-wise or population alignment approaches. First we extend our unified statistical framework to incorporate such techniques, and then we present a new method to attack this challenging task. Our framework, stochastic congealing, provides a computationally efficient formulation for both uni- and multi-modal population registration of 3D data volumes. Through a wide variety of experiments, we show its advantages and carefully validate its results.

Briefly, the main contributions described in this work can be summarized by the introduction of: a unified statistical framework to compare pair-wise registration objective functions; a novel pair-wise registration objective function that incorporates information form both previously aligned data sets and the current observations; an extension to the unified statistical framework to include group-wise registration analysis; and a novel group-wise registration framework that is suited for aligning

multi-modal gray-scale data sets in a computationally efficient manner.

7.2 Future Research Directions

We have been inspired by the success of our group-wise registration framework, and we believe that there are several applications that could build on it or further enhance it. Therefore, in this section, we propose some ideas for future research projects directly related to the idea of stochastic congealing.

7.2.1 Parallel Implementation of Non-rigid Warps

We have discussed the implementation of free-form deformations under the congealing framework in Chapter 6. Through our experiments, we realized that a useful implementation of such a dense field registration option is computationally very expensive, even when run in a multi-resolution fashion. In order to remedy that limitation, we propose a parallelized implementation of the non-linear deformation. Similar efforts, in the case of pair-wise registration, have already proved to be encouraging [62], thus we believe that our algorithm could also benefit from it. More specifically, we note that our objective function, the sum of voxel-wise entropy metric, is well-suited to be parallelized, as the computation of the one-dimensional entropy measures is very localized. Thus subdividing the input volumes into smaller components (with overlapping areas at the division margins) could allow for the computation of a free-form deformation field from small individual regions.

7.2.2 Bias Removal and Spatial Alignment

Recently, our collaborators have been applying the congealing framework to bias removal of MRI images [35, 36]. Instead of optimizing the alignment criterion over transformation components, they recover bias field components that describe non-uniform intensity inhomogeneities in the input images. At the moment, the bias removal mechanism has been applied to input images that are approximately aligned. However, no explicit spatial normalization is required. As the bias removal results are promising, even with respect to slightly misaligned data sets, we are interested in incorporating our spatial alignment technique in their framework. We believe that the recovery of bias components could largely benefit from such an addition. We foresee two directions that could be pursued in order to combine the two distinct congealing mechanisms. First, a sequential implementation of bias removal and spatial alignment could be executed; second, a simultaneous implementation could jointly optimize for both the bias and the spatial transformation unknowns.

7.2.3 Diffusion Imaging Studies

A parallelized and thus computationally affordable computation of the free-form deformation fields could also open the possibility for facilitating the registration of higher

dimensional (non-scalar valued) input images. We have considered the alignment of diffusion tensor images, where instead of a scalar valued component, there is a 3x3 symmetric positive definite matrix that characterizes the image properties at each voxel.

While structural MRI acquisitions provide greatly detailed information about soft tissue types in general, they image the white matter regions as almost homogeneous entities. Thus, in most of the currently existing applications, after the cortical layers are matched and registered, the white matter region is only interpolated. The resulting warp is often not the most desirable one as the deformation does not exploit essential anatomical information about the complex structures in the white matter.

A relatively new imaging modality called diffusion weighted MRI (DW-MRI) has a lot of potential in providing additional information about white matter tracts and the white matter in general. Diffusion weighted images measure the amount of water diffusion along particular directions in the imaged tissues. In white matter, the motion of water molecules is restricted by the walls of the fiber bundles, so the diffusion information can allow for the localization / characterization of the white matter structures. One way to encode the diffusion characteristics is to associate diffusion tensors with each voxel location creating diffusion tensor images (DTI). These images could play a major role in enhancing the construction of white matter warps in the case of inter-subject registration algorithms.

A congealing-based group-wise registration could facilitate significant statistical studies in this domain, too. One such application could be the identification of subtle white matter abnormalities in schizophrenia and other diseases.

7.2.4 Surface-based Registration

Surface-based registration methods are popular in the functional imaging literature [68, 18, 19, 75]. They, instead of working directly in the 3D data coordinate system, extract the surface of the imaged object and define a framework to align these lower-dimensional geometries. This choice is justified by the fact that when head acquisitions are considered, much of the interesting information lies on the surface of the cortex and careful alignment of the gyri and sulci is extremely important. For group-wise studies, when multiple image data sets are to be compared, we believe that implementing the congealing framework could be an advantageous option. In a more ambitious plan, we imagine combining surface-based structural and volumetric functional registration in the same framework.

Appendix A

Maximum Likelihood and Information Theory

In this section we demonstrate how the relationship between the Maximum Likelihood (ML) formulation and information theoretic quantities can be obtained as presented in (Eq. (3.8)), which we repeat here. (Another source of detailed explanation of this formulation can be found in [77].) As a reminder, that equality stated that the finite sample expectation of the likelihood function can be expressed by the sum of two information theoretic entities, a KL-divergence and an entropy metric:

$$E_{p_S}\left[\mathcal{L}_M\left(\mathcal{Y}_S\right)\right] = -\left[D\left(p_S||p_M\right) + H\left(p_S\right)\right],$$

where p_S and p_M indicate the source and model distributions respectively.

We defined the notion of KL divergence in Chapter 3. This measure is a nonnegative quantity, which measures the difference between two probability distributions. Given two probability distributions p and q of the discrete random variable X,

$$D(p||q) = E_p\left[\log\left(\frac{p(X)}{q(X)}\right)\right] = \sum_x p(x)\ \log\frac{p(x)}{q(x)}\ .$$

Further manipulating that definition and applying the definition of the Shannon entropy ($H(X) \equiv -\sum_x p(x)\log p(x)$) results in

$$\begin{aligned} D(p||q) &= \sum_x p(x)\ \log\frac{p(x)}{q(x)} \\ &= \sum_x p(x)\ \log p(x) - \sum_x p(x)\ \log q(x) \\ &= -H(p) - \sum_x p(x)\ \log q(x) \\ &= -H(p) - E_p[log(q(X))]. \end{aligned}$$

And thus

$$E_p[log(q(X))] = -[D(p||q) + H(p)]. \tag{A.1}$$

In Eq.(3.7) of Chapter 3, we defined the normalized likelihood criterion

$$\mathcal{L}_M(\mathcal{Y}_S) = \frac{1}{N}\sum_i \log\left(p\left([u, v_S]_i; M\right)\right). \tag{A.2}$$

If we now take the finite sample expectation of both sides of the expression in Eq.(A.2) with respect to source density p_S where the examined samples are *i.i.d.* distributed, we get

$$E_{p_S}\left[\mathcal{L}_M(\mathcal{Y}_S)\right] = E_{p_S}[log(p_M)]. \tag{A.3}$$

Therefore, using Eq.(A.1), we can establish that are original claim is true, that is:

$$E_{p_S}\left[\mathcal{L}_M(\mathcal{Y}_S)\right] = -[D(p_S||p_M) + H(p_S)]. \tag{A.4}$$

Appendix B

Optimality of Mutual Information

In this appendix, we propose a theoretical proof for the global optimality of the mutual information (MI) registration metric. Although MI has been one of the most popular objective functions in the multi-modal registration literature, the existence of its global maxima about the point of correct registration has been only been observed and exploited empirically. To our knowledge, no sets of conditions have been previously established such that this global optimality criterion could be rigorously proved.

In order to prove our claim, we use the following latent anatomy variable model:

$$p(u, v, l) = p_l(l_1, \cdots, l_N) \prod_i p_{u|l}(u_i|l_i)\, p_{v|l}(v_i|l_i),$$

where the sets $\{u_1, \cdots, u_N\}$ and $\{v_1, \cdots, v_N\}$ represent observations (e.g. pixels or voxels) of two different image modalities at corresponding coordinate system locations and $\{l_1, \cdots, l_N\}$ are a set of latent variables which describe tissue properties (e.g. label types). The model simply asserts the independence of the observations *conditioned* on the latent variables. It does not specify the joint properties of $\{l_1, \cdots, l_N\}$, though a partial or a full description of the these relationships could also be incorporated. A graphical diagram[1] depicting the latent anatomy variable model is shown in Fig. B-1. We selected such a model for our analysis as it represents a sufficient framework with a minimal number of assumptions about the image formation procedure. It also has two notable consequences. First, spatial dependencies in the observations arise directly from known or assumed spatial dependencies in the latent variables. Second, bounds on the spatial dependencies (modulo the unknown transformation) can be *estimated* from the individual imaging modalities. In particular, it is easily derived that the mutual information functions of induced images (for example, MR or CT) lower bound that of the underlying latent anatomy (the segmented image labels or the imaged anatomy); and the mutual information values for the pairs of corresponding

[1] A similar representation incorporating voxel positions has been recently introduced for elastic image registration via conditional probability computations [39].

image elements is always greater than or equal to that of non-corresponding ones.

$$I(u_j;u_k), I(v_j;v_k) \leq I(l_j;l_k) \text{ and} \quad \text{(B.1)}$$
$$I(u_j;v_j) \geq I(u_j;v_k) \ \forall\, j,k = 1,...,N. \quad \text{(B.2)}$$

The inequalities in B.1 and B.2 can be derived from the Data Processing Inequality theorem [14]. We briefly provide a proof corresponding to both of them in the following.

If X, Y and Z are random variables forming a Markov chain ($X \rightarrow Y \rightarrow Z$), then $I(X;Y) \geq I(X;Z)$, i.e. no processing of Y can increase the information that Y contains about X.

Proof I: The relationship between the random variables appearing in inequality (B.1), $v_j \leftarrow l_j - l_k \rightarrow v_k$ (see Fig. B-1), can be rewritten in two different forms using Bayes rule: $v_j \leftarrow l_j \leftarrow l_k \leftarrow v_k$ and $v_j \rightarrow l_j \rightarrow l_k \rightarrow v_k$. Using these formulations and applying the Data Processing Inequality theorem, we obtain:

$$I(v_k;l_k) \geq I(v_k;l_j) \geq I(v_k;v_j) \text{ and } I(l_k;l_j) \geq I(l_k;v_j)$$
$$I(v_j;l_j) \geq I(v_j;l_k) \geq I(v_j;v_k) \text{ and } I(l_j;l_k) \geq I(l_j;v_k)$$

Given $I(X;Y) = I(Y;X)$, we can establish $I(l_j;l_k) \geq I(v_j;v_k)$ for all j and k.

Proof II: In a similar manner, we can obtain the following inequalities for u_j, v_j, l_j, l_k, v_k (see again Fig. B-1):

$$I(v_j;l_j) \geq I(u_j;v_j) \text{ and } I(v_k;l_k) \geq I(v_k;l_j) \geq I(v_k;u_j).$$

Applying Bayes rule, we can establish the following relationships: $v_j \leftarrow l_j \leftarrow l_k \leftarrow u_k$ and $v_j \leftarrow l_j \leftarrow u_j$. As we assume that $I(v_k;l_k) = I(v_j;l_j)$, we need to consider two scenarios: (a) $l_k \rightarrow l_j$ indicates a lossless relationship and (b) $l_k \rightarrow l_j$ indicates a lossy connection. In the former case, $I(u_j;v_k) = I(u_j;v_j)$, and in the latter $I(u_j;v_k) < I(u_j;v_j)$. Therefore, we can conclude that $I(u_j;v_j) \geq I(u_j;v_k)$, which was stated in inequality (B.2).

With these inequalities, we guarantee global extrema for the MI objective function. More specifically, the inequalities in (B.1) and (B.2) show that under the latent variable model (which provides a sufficient condition for our statement), MI as an objective criterion is guaranteed to have a global maximum about the point of correct registration.

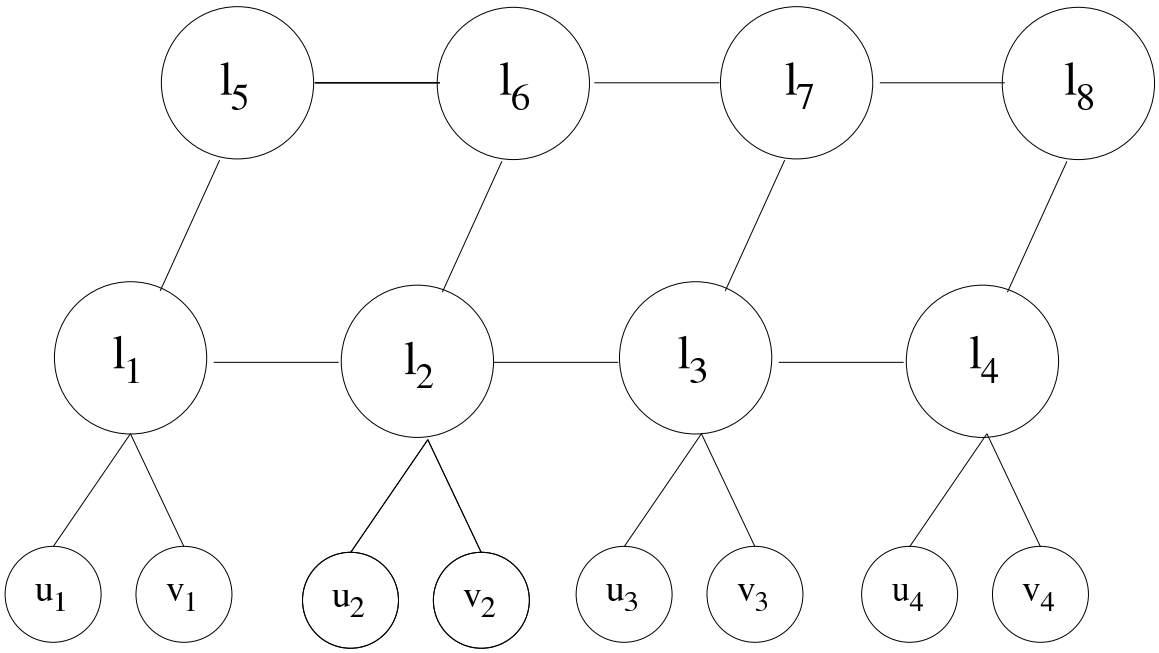


Figure B-1: Example of a latent anatomy model: $\{u_i, v_i\}$ is a correctly aligned voxel pair corresponding to l_i (label/anatomy) at a particular coordinate location; e.g.: pixel, voxel. The connection between the label points is not specified explicitly, the edges connecting l_i's in this figure are indicated just to provide a basic spatial structure to the graph.

Bibliography

[1] J. Beirlant, E.J. Dudewicz, L. Gyorfi, and E.C. van der Meulen. Nonparametric entropy estimation: an overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39, June 1997.

[2] K.K. Bhatia, J.V. Hajnal, B.K. Puri, A.D. Edwards, and D. Rueckert. Consistent groupwise non-rigid registration for atlas construction. In *ISBI*, pages 908–911, 2004.

[3] J.L. Boes and C.R. Meyer. Multi-variate mutual information for registration. In C. Taylor and A. Colchester, editors, *MICCAI*, volume 1679 of *LNCS*, pages 606–612. Springer, 1999.

[4] Morten Bro-Nielsen and Claus Gramkow. Fast fluid registration of medical images. In *Proceedings of the 4th International Conference on Visualization in Biomedical Computing*, pages 267–276, London, UK, 1996. Springer-Verlag.

[5] Lisa Gottesfeld Brown. A survey of image registration techniques. *ACM Computing Surveys*, 24(4):325–376, 1992.

[6] P. Cachier, X. Pennec, and N. Ayache. Fast non-rigid matching by gradient descent : study and improvements of the "demons" algorithm. Technical Report 3706, INRIA, 1999.

[7] D.W. Chen. Oxygen tank becomes fatal missile in hospital. *New York Times*, July 31 2001.

[8] Z.H. Cho, J.P. Jones, and M. Singh. *Foundations of Medical Imaging*. John Wiley & Sons, Inc., 1993.

[9] G.E. Christensen, P. Yin, M.W. Vannier, K.S.C. Chao, J.L. Dempsey, and J.F. Williamson. Large-deformation image registration using fluid landmarks. In *Image Analysis and Interpretation*, pages 269 –273, 2000.

[10] A.C.S. Chung, W.M.W. Wells III, A. Norbash, and W.E.L. Grimson. Multimodal Image Registration by Minimizing Kullback-Leibler Distance. In *MICCAI*, volume 2 of *LNCS*, pages 525–532. Springer, 2002.

[11] O. Clatz, M. Sermesant, P.Y. Bondiau, H. Delingette, S.K. Warfield, G. Malandain, and N. Ayache. Realistic simulation of the 3d growth of brain tumors in mr images coupling diffusion with mass effect. *IEEE Transactions on Medical Imaging*, 24(10):1334–1346, October 2005.

[12] A. Collignon, D. Vandermuelen, P. Suetens, and G. Marchal. 3d multi-modality medical image registration using feature space clustering. In N. Ayache, editor, *Computer Vision, Virtual Reality and Robotics in Medicine*, pages 195–204. Springer, 1995.

[13] D.L. Collins. *3D Model-Based Segmentation of Individual Brain Structures from Magnetic Resonance Imaging Data*. PhD thesis, McGill University, Montreal, Canada, 1994.

[14] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., New York, 1991.

[15] E. D'Agostino, F. Maes, D. Vandermeulen, and Suetens. P. A viscous fluid model for multimodal non-rigid image registration using mutual information. *Medical Image Analysis*, 7:565–575, 2003.

[16] M.H. DeGroot and M.J. Schervish. *Probability and Statistics*. Addison Wesley, 1991.

[17] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, Inc., New York, 1973.

[18] B. Fischl, M.I. Sereno, and A.M. Dale. Cortical surface-based analysis ii: Inflation, flattening, and a surface-based coordinate system. *Neuroimage*, 9(2):195–207, 1999.

[19] B. Fischl, M.I. Sereno, R.B.H. Tootell, and A.M. Dale. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human Brain Mapping*, 8(4):272–284, 1999.

[20] R. Gan and A.C.S. Chung. Multi-dimensional mutual information based robust image registration using maximum distance-gradient magnitude. In Christensen and Sonka, editors, *IPMI*, LNCS, pages 210–221. Springer, 2005.

[21] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. CRC Press LLC, Boca Raton, FL, 2003.

[22] P. Good. *Permutation Tests*. Springer-Verlag, New York, 1994.

[23] P.D. Grunwald, I.J. Myung, and M.A. Pitt. *Advances in Minimum Description Length:Theory and Applications*. The MIT Press, 2005.

[24] A. Guimond, J. Meunier, and Thirion J.-P. Average brain models: A convergence study. Technical Report 3731, INRIA, July 1999.

[25] C. Gütter, C. Xu, F. Sauer, and J. Hornegger. Learning based non-rigid multimodal image registration using kullback-leibler divergence. In G. Gerig and J. Duncan, editors, *MICCAI*, volume 2 of *LNCS*, pages 255–263. Springer, 2005.

[26] J.V. Hajnal, D.L.G. Hill, and D.J Hawkes. *Medical Image Registration*. CRC Press, 2001.

[27] A. Hero, B. Ma, O. Michel, and J Gorman. Applications of entropic spanning graphs. *IEEE Signal Processing Magazine*, 19(5):85–95, Sept 2002.

[28] J.P. Hornak. The basics of mri, (http://www.cis.rit.edu/htbooks/mri/index.html).

[29] A. Ihler, J. Fisher, and A. Willsky. Nonparametric hypothesis tests for statistical dependency. *IEEE Transactions on Signal Processing*, 52(8), August 2004.

[30] H.J. Johnson and G. E. Christensen. Consistent landmark and intensity-based image registration. *IEEE Transactions on Medical Imaging*, 21(5):450–461, May 2002.

[31] S. Joshi, B. Davis, M. Jomier, and G. Gerig. Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage*, 23:S151–S160, September 2004.

[32] Kullback and Solomon. *Information Theory and Statistics*. John Wiley & Sons, Inc, New York, 1959.

[33] E. Learned-Miller. Hyperspacings and the estimation of information theoretic quantities. Technical Report 04-104, UMass Amherst, 2004.

[34] E. Learned-Miller. Data driven image models through continuous joint alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):236–250, January 2006.

[35] E. Learned-Miller and P. Ahammad. Joint mri bias removal using entropy minimization across images. In L.K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 761–768. MIT Press, Cambridge, MA, 2005.

[36] E. Learned-Miller and V. Jain. Many heads are better than one: Jointly removing bias from multiple mrs using nonparametric maximum likelihood. In G.E. Christensen and M. Sonka, editors, *Proceedings of IPMI*, volume 3565 of *LNCS*, pages 615–626. Springer, July 2005.

[37] M. Leventon and W.E.L. Grimson. Multi-modal Volume Registration Using Joint Intensity Distributions. In *MICCAI*, LNCS, pages 1057–1066. Springer, 1998.

[38] P. Lorenzen, B. Davis, and G. Gerig. Multi-class posterior atlas formation via unbiased kullback-leibler template estimation. In *MICCAI*, volume 3216 of *LNCS*, pages 95–102, September 2004.

[39] A.M.C. Machado, M.F.M. Campos, and J.C. Gee. Bayesian model for intensity mapping in magnetic resonance image registration. *Journal of Electronic Imaging*, 12(1):31–39, Jan 2003.

[40] A. Macovski. *Medical Imaging Systems*. Prentice Hall, Inc., Upper Saddle River, NJ, 1983.

[41] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging*, 16(2):187–198, 1997.

[42] J. Maintz and M Viergever. An overview of medical image registration methods, 1996.

[43] J. Maintz and M. Viergever. A survey of medical image registration. *Medical Image Analysis*, 2(1):1–36, 1998.

[44] S. Marsland, C. Twining, and C. Taylor. Groupwise non-rigid registration using polyharmonic clamped-plate splines. In *MICCAI*, volume 2879 of *LNCS*, pages 771–779. Springer-Verlag, November 2003.

[45] S. Marsland and C. J. Twining. Constructing diffeomorphic representations for the groupwise analysis of nonrigid registrations of medical images. *IEEE Transactions on Medical Imaging*, 23(8):1006–1020, August 2004.

[46] A. Mewes, P. Hüppi, H. Als, F.J. Rybicki, T. Inder, R.V. Mulkern, R.L. Robertson, M.J. Rivkin, and S.K. Warfield. Regional brain development in serial mri of low-risk preterm infants. *Pediatrics - to appear*, 2005.

[47] E. Miller, N. Matsakis, and P. Viola. Learning from one example through shared densities on transforms. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 464–471, 2000.

[48] E.G. Miller. *Learning from One Example in Machine Vision by Sharing Probability Densities*. PhD thesis, Massachusetts Institute of Technology, February 2002.

[49] H.F. Neemuchwala and A.O Hero. Entropic graphs for registration. In R.S Blum and Z. Liu, editors, *Multi-sensor image fusion and its applications*. Marcel-Dekker, 2004.

[50] H.J. Park, M. Kubicki, M.E. Shenton, A. Guimond, R.W McCarley, S.E. Maier, R. Kikinis, F.A. Jolesz, and C.F. Westin. Spatial normalization of diffusion tensor mri using multiple channels. *NeuroImage*, 20:1995–2009, 2003.

[51] J.P.W. Pluim, J.B.A. Maintz, and M.A. Viergever. Image registration by maximization of combined mutual information and gradient information. In *MICCAI*, LNCS, pages 567–578. Springer, 2000.

[52] J.P.W. Pluim, J.B.A. Maintz, and M.A. Viergever. Mutual-information-based registration of medical images: a survey. *IEEE Transactions on Medical Imaging*, 22(8):986–1004, 2003.

[53] S. Robbins, A.C. Evans, D.L. Collins, and S. Whitesides. Tuning and comparing spatial normalization methods. *Medical Image Analysis*, pages 311–323, 2004.

[54] A. Roche, G. Malandain, and X. Pennec ad N. Ayache. The correlation ratio as a new similarity measure for multimodal image registration. In *MICCAI*, volume 1496 of *LNCS*, pages 1115–1124. Springer, 1998.

[55] A. Roche, G. Malandain, and N Ayache. Unifying maximum likelihood approaches in medical image registration. *International Journal of Imaging Systems and Technology*, 11(7180):71–80, 2000.

[56] T. Rohlfing, C. Maurer, D. Bluemke, and M. Jacobs. Mr breast images using free-form deformation with an incompressibility constraint. *IEEE Transactions on Medical Imaging*, 22(6):730–741, 2003.

[57] D. Rueckert, S.I. Sonoda, C. Hayes, D.L.G. Hill, M.O Leach, and D.J. Hawkes. Nonrigid registration using free-form deformations: Applications to breast mr images. In *IEEE Transactions on Medical Imaging*, volume 18(8), pages 712–721, 1999.

[58] M.R. Sabuncu and P.J. Ramadge. Gradient based optimization of an emst registration function. In *Proceedings of IEEE Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 253–256, March 2005.

[59] M.R. Sabuncu and P.J. Ramadge. Graph theoretic image registration using prior examples. In *Proceedings of European Signal Processing Conference*, September 2005.

[60] F. Segonne, A.M. Dale, E. Busa, D. Glessner, D. Salat, H.K. Hahn, and B. Fischl. A hybrid approach to the skull stripping problem in mri. *Neuroimage*, 22(3):1060–1075, 2004.

[61] S. Soman, A.C.S. Chung, W.E.L. Grimson, and W. Wells. Rigid registration of echoplanar and conventional magnetic resonance images by minimizing the kullback-leibler distance. In *WBIR*, pages 181–190, 2003.

[62] R. Stefanescu. *Parallel nonlinear registration of medical images with a priori information on anatomy and pathology*. PhD thesis, Universite de Nice – Sophia-Antipolis, March 2005.

[63] C. Studholme and V. Cardenas. A template-free approach to volumetric spatial normalization of brain anatomy. *Pattern Recognition Letters*, 25(10):1191–1202, July 2004.

[64] C. Studholme, D.L.G. Hill, and D.J. Hawkes. Incorporating connected region labelling into automated image registration using mutual information. In *Proc. Mathematical Methods in Biomedical Image Analysis*, pages 23–31, 1996.

[65] C. Studholme, D.L.G. Hill, and D.J. Hawkes. An overlap invariant entropy measure of 3d medical image alignment. *Pattern Recognition*, 32(1):71–86, 1999.

[66] T.S.Y. Tang, R.E. Ellis, and G. Fichtinger. Fiducial registration from a single x-ray image; a new technique for fluoroscopic guidance and radiotherapy. In *MICCAI*, LNCS, pages 502–511. Springer, 2000.

[67] J.-P. Thirion. Non-rigid matching using demons. In *CVPR, San Francisco*, 1996.

[68] P.M. Thompson and A.W. Toga. A surface-based technique for warping 3-dimensional images of the brain. *IEEE Transactions on Medical Imaging*, 15(4):1–16, August 1996.

[69] P.M. Thompson, R.P. Woods, M.S. Mega, and A.W. Toga. Mathematical/computational challenges in creating deformable and probabilistic atlases of the human brain. In *Human Brain Mapping*, volume 9 (2), pages 81–92, February 2000.

[70] S. Timoner. *Compact Representations for Fast Nonrigid Registration of Medical Images*. PhD thesis, Massachusetts Institute of Technology, July 2003.

[71] M. Toews, D. L. Collins, and T. Arbel. Maximum a posteriori local histogram estimation for image registration. In *MICCAI*, volume 2 of *LNCS*, pages 163–170. Springer, 2005.

[72] Harry L. Van Trees. *Detection, Estimation, and Modulation Theory: Part 1*. John Wiley & Sons, Inc., 1992.

[73] C. Twining and C. Marsland, S. Taylor. Groupwise non-rigid registration: The minimum description length approach. In *Proceedings of BMVC*, 2004.

[74] C.J. Twining, T. Cootes, S. Marsland, V. Petrovic, R. Schestowitz, and C.J. Taylor. A unified information-theoretic approach to groupwise non-rigid registration and model building. In G.E. Christensen and M. Sonka, editors, *Proceedings of IPMI*, volume 3565 of *LNCS*, pages 1–14. Springer, July 2005.

[75] D.C. Van Essen, H.A. Drury, S. Joshi, and M.I. Miller. Functional and structural mapping of human cerebral cortex: Solutions are in the surfaces. In *National Academy of Sciences*, volume 95, pages 788–795, 1998.

[76] O. Vasicek. A test for normality based on sample entropy. *Journal of the Royal Statistical Society, Series B*, 38(1):54–59, 1976.

[77] P.A. Viola. *Alignment by Maximization of Mutual Information*. PhD thesis, Massachusetts Institute of Technology, June 1995.

[78] S. Warfield. Fast knn classification for multichannel image data. *Pattern Recognition Letters*, 17(7):713–721, June 1996.

[79] S. Warfield, J. Rexilius, P. Huppi, T. Inder, E. Miller, W. Wells, G. Zientara, F. Jolesz, and R. Kikinis. A binary entropy measure to assess nonrigid registration algorithms. In *MICCAI*, LNCS, pages 266–274. Springer, October 2001.

[80] S.K. Warfield, K.H. Zou, and W.M. Wells. Simultaneous truth and performance level estimation (staple): An algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004.

[81] S. K. Warfiled, M. Kaus, F. A. Jolesz, and R Kikinis. Adaptive, template moderated, spatially varying statistical classification. *Medical Image Analysis*, 4(1):43–55, 2000.

[82] J. Weese, G.P. Penney, P. Desmedt, T.M. Buzug, D.L.G Hill, and D.J. Hawkes. Voxel-based 2-d/3-d registration of fluoroscopy images and ct scans for image-guided surgery. *IEEE Transactions on Information Technology in Biomedicine*, 1(4):284–293, December 1997.

[83] N. Weisenfeld, A.U.J. Mewes, and S.K. Warfield. Segmentation of newborn brain mri. In *IEEE ISBI*, April 2006.

[84] W.M. Wells III, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis. Multi-modal volume registration by maximization of mutual information. *Medical Image Analysis*, 1:35–52, 1996.

[85] W.M. Wells III, P. Viola, and R. Kikinis. Multi-modal Volume Registration by Maximization of Mutual Information [medical imaging]. In *Proceedings of 2nd International Symposium on Medical Robotics and Computer Assisted Surgery*, pages 358,55–62, 1995.

[86] J.B. West, J.M. Fitzpatrick, M.Y. Wang, B.M. Dawant, C.R. Maurer, R.M. Kessler, and R.J. Maciunas. Retrospective intermodality registration techniques: Surface-based versus volume-based. *IEEE Transactions on Medical Imaging*, 18(2):144–150, 1999.

[87] http://en.wikipedia.org.

[88] P. Yushkevich, D. Fritsch, S. Pizer, and E. Chaney. Model-driven determination of 3d patient setup errors in conformal radiotherapy. *Med. Phys.*, 1999.

[89] J. Zhang and A. Rangarajan. Multimodality image registration using an extensible information metric and high dimensional histogramming. In Christensen and Sonka, editors, *IPMI*, LNCS, pages 725–737. Springer, 2005.

[90] Y.M. Zhu and S.M. Cochoff. Likelihood maximization approach to image registration. *IEEE Transactions on Image Processing*, 11(12):1417–1426, 2002.

[91] A.P. Zijdenbos, B. M. Dawant, R.A. Margolin, and A.C. Palmer. Morphometric analysis of white matter lesions in mr images: Method and validation. *IEEE Trans. Med. Imag.*, 13(4):716–724, 1994.

[92] B. Zitová and J. Flusser. Image registration methods. *Image and Vision Computing*, 21(11):977–1000, October 2003.

[93] L. Zöllei. 2d-3d rigid-body registration of x-ray fluoroscopy and ct images. Technical Report AITR-2002-001, MIT, 2001.