Entropy-based Image Registration

Mert Rory Sabuncu

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Electrical Engineering

November 2006

Abstract

This thesis investigates the employment of different entropic measures, including Rényi entropy, in the context of image registration. Specifically, we focus on the entropy estimation problem for image registration and provide theoretical and experimental comparisons of two important entropy estimators: the plug-in estimator and minimal entropic graphs. We further develop an image registration framework based on the graph-theoretic estimator. Within this framework, we address practical and theoretical issues such as the incorporation of spatial information, the efficient and fast search of the optimum alignment, and the employment of previously aligned image pairs. These analyses yield fast, robust and accurate multi-modal affine registration algorithms applicable to different medical problems. Next, we investigate the nonrigid registration problem and provide a novel fast entropy-based nonrigid registration algorithm. Finally, we discuss a scientific application, the normalization of the human cerebral cortex based on patterns of functional response, and investigate an algorithm that employs a correlation-based entropic measure.

Acknowledgements

This thesis wouldn't have existed without the support and guidance of my advisor, Peter Ramadge. He has inspired me in his approach to problems, attention to details and demand for perfection, as a researcher, teacher, writer, and mentor. By learning from him, I feel I have become a better academic. I thank my dissertation readers, Sanjeev Kulkarni and Christophe Chefd'hotel, for their timely feedback and insightful comments. Their input was invaluable to me.

I owe many thanks to my "lab mates", Shannon Hughes, Eugene Brevdo and Bryan Conroy. They were the unofficial readers of this thesis and their constructive criticism has improved this document substantially. I need to extend my gratitude to my first, unofficial advisor in grad school, Vitali Zagorodnov. He was the first person to convince me that I am "PhD material."

I would like to acknowledge Jim Williams, Frank Sauer and Chenyang Xu at Siemens Corporate Research for their generous support. I gained invaluable knowledge and experience from my time at SCR and collaboration with Christophe.

My research life in the last year and half has been particularly enjoyable, thanks to the exciting and stimulating collaboration with the Center for the Study of Brain, Mind and Behavior. It was an honor to work closely with Jim Haxby and Ben Singer. I will never forget the birth of "func norm" and the evolution of our results from "promising" to "stunning!"

The five years I spent in this little town called Princeton have been unforgettable thanks to my friends. Here, I thank Eko, Cano, Fato, Ayşen, Seven, Kory, Vitali, and Spanish gentleman on a bike, Alvaro, for their support and distraction.

I thank my parents, Mustafa and Maggie, and my little brother Metin for their constant love and support. They continue to be inspirations of mine at every stage of my life. Finally, thanks to my Filiz, for being beside me - everything in its right place...

To the Sabuncu and McGaughey Families.

Contents

Chapter 1

Introduction

This thesis deals with the fundamental problem of image (signal) alignment and investigates different techniques to solve the problem using ideas that reside on the boundary of image processing, computer vision and information theory. Parallel to recent trends in computer vision, e.g. [97], in our approach, we look at the alignment problem from a stochastic viewpoint and employ rigorous results from the information, probability and graph theory literatures to design practical and useful algorithms.

We view this thesis as a continuation of the exploration of fast, efficient, robust and flexible algorithms intended for signal alignment. We are motivated to study these issues since we believe that no matter how advanced our computational technology becomes, mankind will always be faced with challenging applications that are constrained by the limitation of physical resources, such as space, time and bandwidth.

In this thesis, alignment is typically performed on functions defined in a two- or three-dimensional domain, where space is the independent variable. As commonly done in the literature, these functions will be called *images*. Moreover, the alignment problem will generally be referred to as *image registration*. As algorithm designers, we are not concerned about the source of the images, which can be a physical sensor, or even a computer model. Also, note that most of the discussed ideas can straightforwardly be applied to other alignment problems in different application domains.

In image processing, often (for example, when comparing or combining the information content of images), we are interested in the relationship between two or more images. The analysis of this relationship usually becomes tractable once a *correspondence* is set up between the images. Image registration is the task of setting up this correspondence.

Image registration shows up in a rich range of application domains, such as medical image analysis (e.g. diagnosis), neuroscience (e.g. brain mapping), computer vision (e.g. stereo image matching for shape recovery), astrophysics (e.g. the alignment of images from different frequencies), military applications (e.g. target recognition), etc. For a detailed overview of these applications, see [33, 52]. The definition of correspondence varies across disciplines and even across individual applications. Thus, a proposition of a universal image registration algorithm is practically impossible. Yet, in this thesis our goal is to keep the approach as general as possible, so that discussed techniques can be tailored towards the specifics of a particular application.

This thesis will frequently be using medical images as examples to illustrate and support ideas. Due to the rapid advancement of imaging technologies in recent decades, we are enjoying a widespread availability of medical imaging devices, such as the Magnetic Resonance (MR), Ultrasound (US), X-ray, Positron Emission Tomography (PET), etc. The images acquired by these devices display anatomical (structural) and/or functional information about the imaged organ or body part. Fusion and/or comparison of the information content of different images is usually achieved based on image registration. This information is then used for various purposes, such as detecting pathological changes, treatment verification, early diagnosis, scientific research, etc.

Image registration is particularly difficult when images are obtained through different sensor types (*multi-modal registration*) and/or when complex (e.g. nonlinear) geometric transformations are required to relate the images, e.g. when registering images of different human brains (*multi-subject registration*). Much of this thesis investigates the idea of employing *information theory* to solve this difficult problem. Most of the ideas are contributions to the school of thought inspired by the influential papers of Collignon et al. [50] and Viola and Wells [97]. The advantage of this approach is that it provides sufficient flexibility for capturing the underlying relationship between the images. Note that this relationship is typically complex and unknown in a multi-modal setting. This strength of the approach has lead to the design of many robust and accurate multi-modal registration algorithms. For a detailed survey of these algorithms the reader is referred to [68].

Although, the information-theoretic approach has yielded many useful registration algorithms, it has some drawbacks. Most importantly, the computation of these similarity measures is *typically* slower than competing simpler measures, such as the L^2 norm or correlation. This becomes critical when time is an important constraint of the application. An important contribution of this thesis is a detailed analysis of this problem and several proposed approaches to design fast algorithms.

Moreover, most of the current information-theoretic registration algorithms rely on the strong assumption that pixel intensity values are independent and identically distributed samples of a random variable. This causes the algorithm to discard *spatial information* in the images. As a result, registration accuracy may suffer. We include a brief discussion of this issue and propose methods to alleviate the problem.

As presented in this thesis, stochastic models of the registration problem yield different entropic measures that quantify the quality of alignment. These measures, in theory, are functionals of underlying probability densities. In practice, however, the algorithm only has access to a finite number of samples. Thus, it has to *estimate* the entropic measure (see [3]). A major contribution of this thesis is the analysis and comparison of different entropy estimators intended for image registration. Particularly, we focus on two techniques: *plug-in* estimators and *minimal entropic graphs*. Plug-in estimators are more straightforward to implement and readily provide gradient information, which can be used for the efficient search of the optimum transformation. Thus, they have been popularly employed for image registration. Entropic graphs, on the other hand, have only recently been applied to registration [60]. We show that despite the fact that the entropic graph estimator yields a non-differentiable measure, it is possible to efficiently compute a descent direction which can then be used for the fast optimization of the alignment measure.

In certain real-world applications, *previously registered image pairs* are available to the algorithm. In a multi-modal application, these pre-registered images contain valuable information about the cross-modality relationship. Another contribution of this thesis is a proposed method to incorporate prior knowledge into the entropic graph based framework. We show that the incorporation of prior knowledge improves robustness against bad initialization but can reduce accuracy due to imperfect prior knowledge. To avoid this, we suggest a simple remedy. Moreover, we propose a method to minimize the computational overhead introduced by training that uses an entropic graph computed off-line over the training samples.

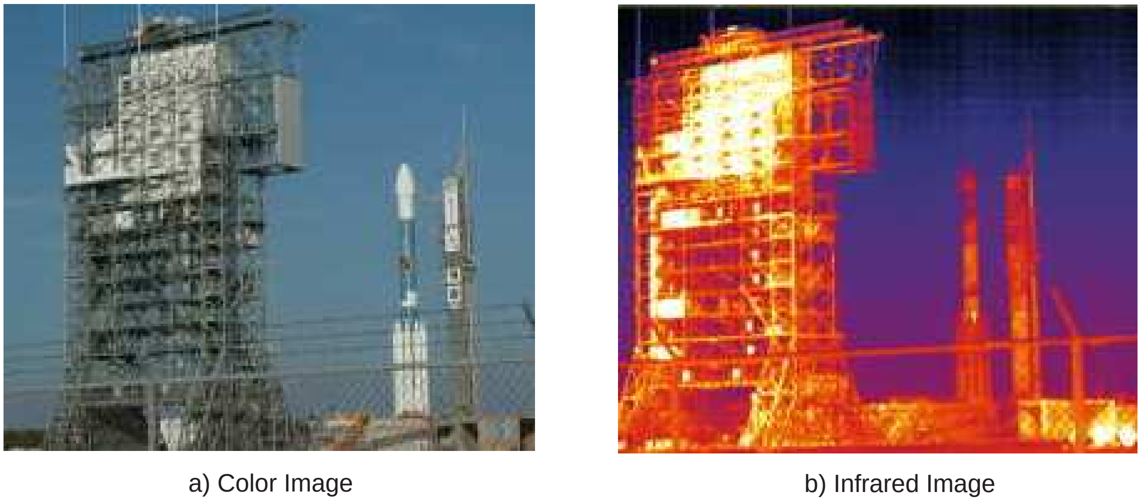
Global transformation models, e.g. rigid-body, are in many cases (e.g. for multi-subject registration) not sufficient to recover the misalignment. In this thesis, we use the broad term of *nonrigid registration* to refer to such problems. Until recently, the multi-modal and nonrigid (registration) literatures had evolved separately, mainly because of the high complexity of the problem [65]. The combination of the two approaches using a two-step algorithm was a common technique. These algorithms typically establish cross-subject/nonrigid alignment using high resolution images of a reference modality and mono-modal registration algorithms that employ simplistic similarity measures, e.g. sum of squared differences of pixel intensity values, and high dimensional geometric transformations. Images of other modalities are then registered with the reference modality image within each subject using low dimensional global geometric transformations. A typical example of this approach is used for functional MRI studies, where cross-subject registration is achieved based on hi-res structural MRI scans (mono-modal, cross-subject, nonrigid registration using the reference modality) and the fMRI data sets are aligned with each subject's structural MRI volume using rigid-body transformations that correct for subject motion during the scan. Note that, this procedure indirectly registers a subject's fMRI data set to another subject's structural and functional MRI volumes.

As new imaging modalities become available, the need for having robust and accurate nonrigid multi-modal registration algorithms to fuse and compare the information content of these modalities increases. Moreover, in many of today's applications, e.g. perfusion studies, cardiac motion analysis, etc., the aforementioned two-stage strategy either is not applicable due to the lack of a reference modality or does not provide satisfactory results. Hence, the requirement for nonrigid multi-modal image registration algorithms. Over the last several years, information-theoretic methods have been applied to this problem. However, for most of these methods, speed is a critical issue because of the computational complexity of the employed similarity measure and/or high-dimensional nature of the optimization problem. In this thesis, we investigate a fast and accurate multi-modal nonrigid registration algorithm that relies on a one-dimensional *"level set entropy"* measure.

1.1 Image Registration: An Overview

In their survey papers, Brown [6] and Maintz and Viergever [52] provide excellent overviews and categorizations of image registration techniques. Most of the ideas presented in this section are based on these surveys.

The variations (sometimes called distortions) across two or more images of the same (or similar) scene can be grouped into two categories: *spatial* (geometric) and *intensity* (valumetric) variations. For example, Figure 1.1 shows a color and an infra-red (IR) image of the same scene taken from slightly different angles (geometric distortion). Pixels in both images that correspond to the same structure contain values in different ranges (valumetric distortion). Image registration attempts to correct for (recover) some of these variations, while preserving others.



a) Color Image

b) Infrared Image

Figure 1.1: A color and IR image of the Spitzer Space Telescope's Delta II rocket obtained from NASA's website (http://solarsystem.nasa.gov/multimedia).

Geometric variations can be divided into two groups: Those we want to correct for and those we don't. A typical example for the first group is variations due to viewpoint changes, i.e., different orientations of the imaging device. On the other hand, we may want to preserve variations due to changes in the scene that are of interest, for example when monitoring tumor growth.

Valumetric variations are due to three main reasons:

- Scene differences: For example, an alien object may be present in one of the images, changing the range of intensity values in the corresponding region.

- Different sensor type: Imaging devices can measure everything from hydrogen density (MRI) to temperature (thermography). Thus, the same physical reality can yield very different representations, which is the case in a multi-modal setting. These representations usually contain correlated information along with complementary, uncorrelated information.

- Different conditions: For example different lighting in conventional cameras, different magnetic field properties in the MRI machine, etc.

When designing a registration algorithm, it is important to identify the variations we want to correct for. The algorithms that this thesis deals with, do not attempt to recover valumetric variations, i.e., changes in the intensity ranges. It is, however, useful to have a good understanding of how intensity values in the different images are related. This knowledge can then be employed to identify and efficiently recover geometric variations of interest.

A typical image registration algorithm consists of three coupled components: *an alignment measure* (also known as similarity measure, registration function, etc.) that quantifies the quality of alignment; *a class of admissible geometric transformations* that can be applied to the image(s), i.e., employed to spatially "warp" the image(s); and *an optimizer* that seeks the transformation that maximizes the similarity as quantified by the alignment measure. Figure 1.2 illustrates these components.

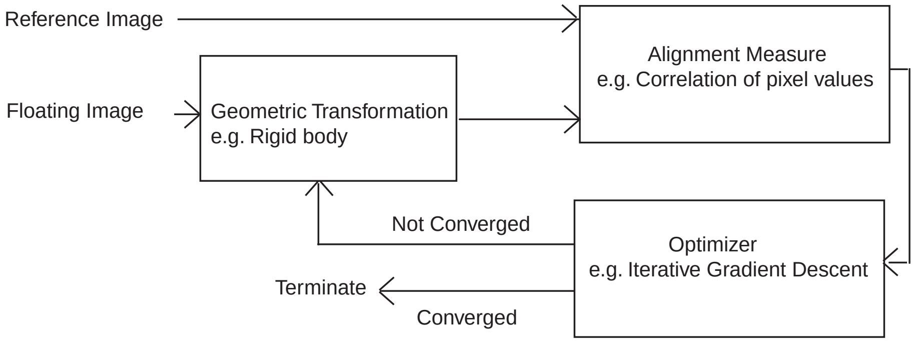


Figure 1.2: A Block diagram that represents a typical image registration algorithm.

Roughly speaking, registration algorithms can be analyzed under two different categorizations: *application-based* and *methodology-based.* Note that there is a strong relationship between these two types of classifications. The application defines alignment (i.e., what we mean by correspondence) and it determines the constraints, e.g. on time and memory. These specify (or, more broadly speaking, narrow down) the choices of algorithms, i.e., the methodology.

From the application's point of view, registration algorithms can be classified based on several criteria. The classifications presented here are partially based on [93]. The criteria and their primary subdivisions are:

- **Modality:** *Mono-modal* refers to the case where all images are obtained from the same imaging sensor type and there are no major differences between the intensity ranges that correspond to the same physical/physiological phenomenon. In a *multi-modal* setting, these ranges can differ drastically. This is typically due to different sensor types.

- **Dimensionality:** This refers to the number of dimensions of the *images.* Historically, images have typically had two spatial dimensions. Today, however, several imaging technologies provide 3D *volumes.* Moreover, some sensors, e.g. functional MRI, provide a video, i.e. a sequence of images. When treating the video as one big data set, time can be thought of as an extra dimension. One convention is to denote time as a 0.5 dimension. This is helpful to clarify some ambiguities, e.g. 3D (three spatial dimensions) versus 2.5D (two spatial dimensions + time). Most of today's applications involve 2D/2D and 3D/3D registration.

- **Speed:** *Offline* refers to applications where time is not an important constraint. *Online* denotes a heavy time constraint, typically indicating real-time applications. An important online example is intra-operative procedures performed within the operating theater. Some scientific applications (e.g. human brain mapping), on the other hand, do not have a heavy time constraint.

- **Subject:** In a medical application, *intra-subject* refers to the task where all images are of the same subject (patient). *Inter-subject* denotes the fact that more than one subject is involved. If an "averaged" template (atlas) is employed, this is typically called *atlas registration*. Inter-subject applications are typically more complex, since correspondence is difficult to identify.

- **Nature of Misalignment:** Geometric misalignment can be attributed to several factors, including different *viewpoints* (orientation) of sensors, *temporal* changes (e.g. Digital Subtraction Angiography: the registration of images before and after radio isotope injections to characterize functionality) and *inherent differences* (e.g. brains of different subjects).

From the methodology point of view, registration algorithms can be classified based on several criteria:

- **Employed Information Content:** In the registration literature, one can identify two trends in the type of information employed. *Landmark* based approaches rely on the definition of landmarks. Alignment is computed based on these landmarks (sets of points, lines or surfaces) only. These landmarks can have a clear physical meaning (e.g. the cortical surface of the human brain [19], fiducial markers visible in all modalities [94], etc.), or they can be of theoretical interest only (e.g. lines, corners, points of high curvature, etc.). In landmark based registration, the set of identified points is sparse compared to the original image content, which allows fast optimization. However, performance of the algorithm heavily depends on the landmark identification. *Image content* based approaches, on the other hand, rely on pixel intensity information. These typically extract features from pixels (e.g. intensity values [97], gradient vectors [67], wavelet coefficients [59], etc.) and compute an alignment based on the set of feature samples. These are usually slower than landmark based algorithms, but have the potential to produce accurate and robust results in contexts where landmarks are difficult to define or determine.

- **Locality of Alignment Measure:** Alignment quality can be measured for the whole image, using *global* measures, e.g. sum of squared differences of all pixel values, or for a neighborhood of a pixel location using *local* measures, e.g. local correlation.

- **Transformation:** Generally speaking, there are two types of geometric transformations: *parametric* models, e.g. rigid-body, affine, spline based, etc., where a small set of parameters determine the transformation and *nonparametric* models (also known as optical flow, dense matching, etc.), where each pixel is allowed to move independently. Note that in the latter case, if there was no restriction on the transformation, an image could be made to look similar to any other image with the same intensity range as the first image. Thus, these methods require regularization to overcome ill-posedness and incorporate prior knowledge about the deformation field.

- **Optimization:** Typically, iterative methods are employed within a multi-resolution pyramid, to speed up convergence. Popular choices of optimizers are: *gradient-descent* and its variants [97], *Powell's method* [50], *Downhill simplex method* and *Levenberg-Marquardt* optimization [51].

In this thesis, we restrict our analysis to *global image content-based* approaches, which provide a general framework and require minimal knowledge about the specifics of the application domain.

1.2 Image Registration: Theory

In this section, we overview the theoretical aspects of an image registration problem.

1.2.1 Problem Definition

Let $U_j(\cdot)$ be in a family $\mathcal{U}_j$ of scalar valued images defined on Ω, a finite subset of $\mathbb{R}^d$, $d \in \mathbb{Z}^+$. For example, all brain MRI volumes may constitute a family, $\mathcal{U}$. The relationship between any two images U_1 and U_2 can be written as:

$$U_1(\cdot) = f(U_2 \circ \Phi)(\cdot) + N(\cdot), \tag{1.1}$$

where $\Phi : \mathbb{R}^d \mapsto \mathbb{R}^d$ is a geometric transformation that models the misalignment that we want to recover, $f : \mathcal{U}_2 \mapsto \mathcal{U}_1$ is a cross-image family mapping that captures valumetric variations and $N : \mathbb{R}^d \mapsto \mathbb{R}$ is some noise, i.e., a scalar image. In this model $(U_1, U_2 \circ \Phi)$ is a (optimally) registered pair of images. The goal of the algorithm is to estimate Φ, by maximizing an alignment measure (or, minimizing a misalignment measure).

Alignment Measure

In Equation (1.1), we can model U_1, U_2, N, and Φ as random variables. For example, U_j can have a uniform distribution on all images in $\mathcal{U}_j$, Φ can be uniformly distributed over some admissible set of rigid-body transformations, and N can be a Gaussian random field.

If f is known, in a maximum likelihood framework, the registration problem can be set up as:

$$\arg\max_{\Phi} p(U_1, U_2|\Phi). \tag{1.2}$$

To compute (1.2), we need to make further modelling assumptions. For example, in a mono-modal setting, we can assume f to be the identity function and the noise to be i.i.d Gaussian. Then, it is easy to show that the log-likelihood function of (1.2) is proportional to:

$$\log p(U_2, U_1|\Phi) \propto -\sum_{\mathbf{x}\in\Omega}(U_1(\mathbf{x}) - U_2(\Phi(\mathbf{x})))^2, \tag{1.3}$$

which is the sum of squared differences, SSD.

Another common alignment measure is the normalized cross-correlation (NCC) [6]. This is based on the assumption that there is a linear relationship (up to some noise) between corresponding pixel intensity values. NCC is defined as:

$$NCC(U_2 \circ \Phi, U_1) = \frac{\sum_{\mathbf{x}} U_2(\Phi(\mathbf{x}))U_1(\mathbf{x})}{\sqrt{\sum_{\mathbf{x}} U_2(\Phi(\mathbf{x}))^2}}. \tag{1.4}$$

This is related to the well-known Pearson's correlation r between corresponding intensity values of the two images:

$$r(U_2 \circ \Phi, U_1) = \frac{\sum_{\mathbf{x}}(U_1(\mathbf{x}) - \mu_1)(U_2(\Phi(\mathbf{x})) - \mu_2)}{(N-1)\sigma_1\sigma_2}, \tag{1.5}$$

where N is the number of pixels, μ_i and σ_i are the mean and variance values of the pixel intensity values in image U_i. Notice that, if we assume μ_2 does not depend on Φ, then minimizing (1.5) over Φ is equivalent to minimizing (1.4).

However, in most applications, we don't know (or even have a good model for) f and thus may have to infer it from the images. A generalized maximum likelihood approach can then be employed, where (1.2) is replaced by:

$$\arg\max_{\Phi}\max_{f} p(U_1, U_2|\Phi, f). \tag{1.6}$$

In the following chapter, we will discuss this issue and motivate information-theoretic measures that can handle unknown cross-image family mappings.

1.2.2 Geometric Transformations

Different transformation models are utilized for various registration applications. Recall that, the geometric transformation attempts to recover the "to-be-corrected" spatial misalignment, e.g. camera/object motion, while some spatial misalignment may want to be preserved, e.g. due to tumor growth. For some applications, e.g. inter-subject registration, the notion of "to-be-corrected" is difficult to define. Thus, it is important to identify the type of misalignment we want to recover, i.e., define the transformation space, based on the specifics of the application.

In general, there are two approaches to define a geometric transformation: using *parametric* models, and using a *dense deformation field*, i.e., in a nonparametric fashion. The first approach employs a small number of parameters to define the warp, whereas the latter method uses a (deformation/motion) vector at each pixel location.

In a parametric transformation model, typically all possible parameter values are treated as being equally likely[1]. In the following, for the sake of compactness, we assume a two-dimensional space, i.e., $\mathbf{x} = (u, v) \in \mathbb{R}^2$ and $\mathbf{x}' = \Phi(\mathbf{x})$. Some commonly used parametric transformation models are:

- **Affine:** In 2D, it is parameterized by six parameters $(a_0, a_1, a_2, b_0, b_1, b_2)$:

$$\begin{aligned} u' &= a_0 u + a_1 v + a_2 \\ v' &= b_0 u + b_1 v + b_2, \end{aligned}$$

  which can map a parallelogram onto a square. This model is defined by three non-collinear corresponding points, preserves straight lines and straight line par-

[1] This is not true for some nonlinear registration algorithms that employ splines, see e.g. [79]

allelism. Rigid-body (rotation and translation) and similarity (rotation, translation and global scale) are special cases.

- **Projective:** It uses eight parameters $(a_0, a_1, a_2, b_0, b_1, b_2, c_1, c_2)$:

  $$\begin{aligned} u' &= \frac{a_0 u + a_1 v + a_2}{1 + c_1 u + c_2 v} \\ v' &= \frac{b_0 u + b_1 v + b_2}{1 + c_1 u + c_2 v}, \end{aligned}$$

  and is commonly used in the pin-hole camera model.

- **Polynomial:** This is a generalization of the affine model and can be expressed as:

  $$\begin{aligned} u' &= \sum_{i=0}^{m} \sum_{j=0}^{m} a_{ij} u^i v^j \\ v' &= \sum_{i=0}^{m} \sum_{j=0}^{m} b_{ij} v^i u^j, \end{aligned}$$

  where the order m determines the "richness" of the transformation.

- **Radial-basis:** This method provides a group of global transformations that can handle local distortions. In general, they can be expressed as:

  $$\begin{aligned} u' &= a_0 + a_1 u + a_2 v + \sum_i c_i g(\|\mathbf{x} - \mathbf{x}_i\|) \\ v' &= a_3 + a_4 u + a_5 v + \sum_i d_i g(\|\mathbf{x} - \mathbf{x}_i\|), \end{aligned}$$

  where $\mathbf{x} = (u, v)$, $\mathbf{x}_i$'s are called control points and g is the radial basis function. Popular choices for g are the thin-plate spline: $g(r) = r^2 \log r$, and B-splines, e.g. [80].

The advantage of parameterized techniques is that the dimensionality of the problem is relatively low and thus robust optimization is possible. However, in some applications it is not clear how to select a natural parameterized transformation space.

In a nonparametric approach each image pixel is transformed independently. One popular technique to impose some regularization on this formulation employs a global objective function that consists of two terms: the alignment measure and an external regularization term that reflects our expectations by penalizing unlikely transformations. Other methods employ a Bayesian approach with a prior distribution model, e.g. Brownian warps [61]. An alternative strategy is an iterative scheme where a "rough" warp field obtained from the gradient of the similarity measure is projected onto a known function space. This projection is done by spatial smoothing [65] and has yielded fast nonrigid registration algorithms [24].

1.2.3 Optimization

Given an alignment measure, $S(U_1, U_2)$, and a family of geometric transformations, $\mathcal{W}$, registration is merely an optimization problem:

$$\Phi^* = \arg\max_{\Phi \in \mathcal{W}} S(U_1, U_2 \circ \Phi). \tag{1.7}$$

Some methods, e.g. Fourier based algorithms [42] that deal with simple transformation spaces (e.g. translation only) and simple alignment measures (e.g. SSD), can solve (1.7) directly. Most methods, on the other hand, do not enjoy a well-behaved, low dimensional objective function. Typically, registration algorithms attempt to solve the optimization using an iterative strategy. For a detailed survey, see [51]. With a parameterized family of transformations, the goal is to search for the optimum set of parameter values. Note that the similarity measure gradient (with respect to transformation parameters) is commonly used to speed up this search.

1.3 Overview of Thesis

The second chapter of this thesis contains a detailed derivation of information-theoretic alignment measures and an overview of different implementations. We provide a discussion of the main advantages and drawbacks of these algorithms from the perspectives of different performance criteria, such as speed, accuracy, robustness, etc. Chapter 3 focuses on the entropy estimation problem and includes a detailed analysis and comparison of two entropy estimators intended for image registration, namely the plug-in estimator and minimal entropic graphs. In Chapter 4, we apply these estimation techniques to the rigid registration problem and introduce a novel graph-theoretic registration framework. Also, in this chapter we continue the analysis of these entropy estimators from the perspective of image registration. The insights provided by this analysis is intended to be a major contribution of this thesis. In Chapter 5, we consider the problem of incorporating prior knowledge about the relationship between two images. Particularly, we focus on the case where the algorithm has access to previously aligned image pairs. We propose a novel alignment measure that utilizes this information to better the algorithm's performance. Chapter 6 includes a discussion of nonrigid registration algorithms and introduces a novel, fast entropy-based registration algorithm. In Chapter 7, we focus on an important scientific application, the functional alignment of the human cerebral cortex, and discuss the application of entropic measures to this problem. Chapter 8 concludes with a discussion of the contributions of this thesis and possible directions for future research. Various ideas discussed in Chapters 1-6 were reported in [81, 82, 83] and [84].

Chapter 2

Information-theoretic Alignment Measures

In the following chapters, we will utilize information-theoretic ideas for image registration. Here, we give some background on these ideas, motivate their employment for alignment and provide an overview of the literature.

Section 2.1 introduces and provides definitions for the notions of entropy, information and uncertainty. Other measures based on the entropy definition, are provided in the following section. Section 2.3 discusses the use of entropic measures to quantify the quality of alignment in image registration. Two derivations based on a maximum likelihood and hypothesis testing framework are included. Section 2.4 provides a discussion of practical issues and highlights the advantages and drawbacks of different implementations.

2.1 Information Entropy and Uncertainty

Claude Shannon's 1948 paper, entitled "A Mathematical Theory of Communication," [85] is widely accepted as the birth of Information Theory. For a detailed treatment of the subject, the reader is referred to more dedicated works, such as [16]. An excellent historical overview is also presented in [95]. Until 1948, little progress had been made to introduce a universal measure for the quantity of information a data source possesses. In [85], Shannon uses probability theory to model information sources, i.e., the data produced by a source is treated as a random variable. The information content, namely (Shannon's) *entropy* of a discrete random variable X that has a probability distribution $\mathbf{p}_X = (p_1, \ldots, p_n)$ is then defined as:

$$H(X) = H(\mathbf{p}_X) \triangleq \sum_{i=1}^{n} p_i \log(1/p_i), \tag{2.1}$$

where $0 \log \infty = 0$ and the base of the logarithm determines the unit, e.g. if base 2 the measure is in *bits*, if it's the natural number e then it's in *nats*, etc. The term $\log 1/p_i$ indicates the amount of *uncertainty* associated with the corresponding outcome. It can also be viewed as the amount of *information* gained by observing that outcome. Thus, entropy is merely a statistical average of uncertainty or information.

Shannon also provides an axiomatic derivation of (2.1): This is the only function of $\mathbf{p}$ that is continuous with $\mathbf{p}$; increases with n; and is additive, i.e., the entropy of two random variables is the sum of the entropy of the first and the entropy of the second given the first. Yet, this derivation is not the key reason that entropy plays a central role in today's information theory. Using (2.1), many information-theoretic results can be derived concisely. For example, it is known that a uniquely decipherable code required for X has a minimum average length bounded by $H(X)$ and $H(X)+1$.

Some properties of (2.1) are:

1. $H(X) \geq 0$ and is equal to zero if and only if X is deterministic.

2. Entropy is the greatest when all samples are equally likely, i.e., $H((p_1, \ldots, p_n)) \leq \log n$.

3. Let X_1 and X_2 be two discrete random variables with joint probability $\mathbf{p}_{X_1,X_2} = \{p_{ij}\}$. Define $H(X_1, X_2) = \sum_{x_1,x_2} p_{ij} \log 1/p_{ij}$. Then $H(X_1, X_2) \leq H(X_1) +$

$H(X_2)$.

4. Entropy doesn't depend on the value of the random variable, but only depends on the distribution. So, for a bijective mapping $f : \Omega_X \rightarrow \Omega_X$, where Ω_X is the domain of X:

$$H(f(X)) = H(X). \tag{2.2}$$

5. Entropy is a concave function of $\mathbf{p}$, i.e.;

$$H(\beta\mathbf{p}_1 + (1-\beta)\mathbf{p}_2) \geq \beta H(\mathbf{p_1}) + (1-\beta)H(\mathbf{p}_2),\ \forall\beta \in [0,1].$$

For a continuous random variable Y that has a probability density $p_Y(\cdot)$, Shannon's *differential entropy* is:

$$H(Y) = H(p_Y) \triangleq -\int p_Y(y)\log p_Y(y)dy. \tag{2.3}$$

An important difference between the discrete and continuous entropies is that, while the discrete entropy is an *absolute* measure of randomness, the differential entropy is a relative measure that depends on the coordinate system. The differential entropy in general can be negative and can achieve arbitrarily small values.

In summary, entropy can be viewed in various ways: a measure of uncertainty in a random event (i.e., a measure of the "randomness" of a random variable), a measure of information obtained by observing a data source, and the dispersion, i.e., scatterdness of a probability distribution.

2.1.1 Entropy of an Image

For a high dimensional discrete random variable $\mathbf{X} = (X_1, \ldots, X_d) \in \mathbb{R}^d$ that has a probability mass function of $p(x_1, \ldots, x_d)$, the entropy formula (2.1) can be extended straightforwardly:

$$H(\mathbf{X}) = \sum_{x_1,\ldots,x_d} p(x_1,\ldots,x_d) \log \frac{1}{p(x_1,\ldots,x_d)}. \tag{2.4}$$

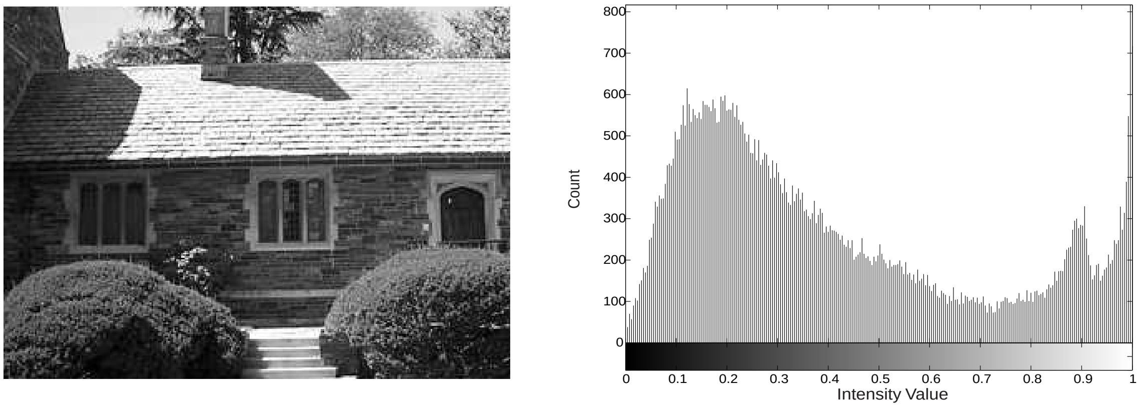
Note that if X_i's are independent and identically distributed with a p.m.f. q for all i, it is easy to see that $H(\mathbf{X}) = d \cdot H(q)$. In information theory, an information source that produces such a random variable is usually called stationary and memoryless. Note that, in a general stationary source, i.e., if X_i's are identically distributed with q; then $H(\mathbf{X}) \leq d \cdot H(q)$. That is the joint random variable cannot contain more information than the sum of the individual information entropies of the components. The upper bound is only achieved when all components are independent.

Similar to Shannon's treatment of the English language in [85], we can analyze images as realizations of random variables. A simple model would assume that each pixel is an i.i.d. realization. Figure 2.1 shows a natural image ("house") and a histogram of the pixel intensity values. The normalized histogram can be an estimate of the underlying probability of pixel intensities, i.e., $p(i) = h_U(i)/N$, where $h_U(i)$ denotes the histogram entry of intensity value i in image U and N is the total number of pixels of U. Using this model, we can compute the entropy of the image as:

$$H(U) = \sum_i h_U(i) \log N/h_U(i), \tag{2.5}$$

which for the image shown in Figure 2.1 is approximately 5.42×10^5 bits.

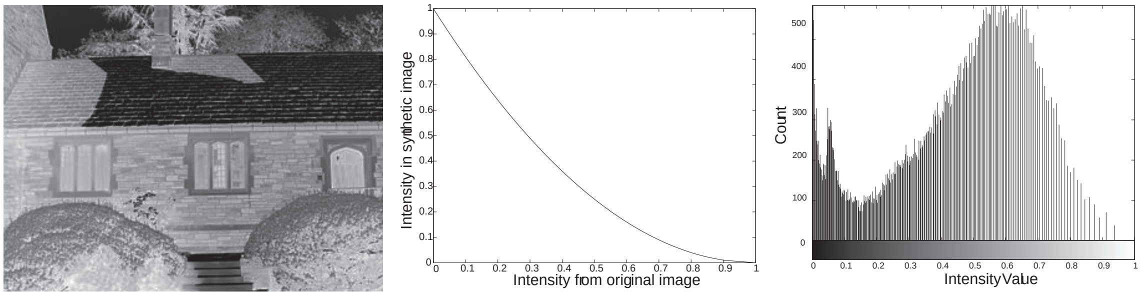
Now, let's apply a bijective mapping to the intensity values of an image. Then, from (2.2) it is easy to see that the entropy of the image does not change. Figure 2.2 shows a synthetic image created by applying a bijective mapping to the "house" image. The corresponding histogram is also shown. Note that, even though the shape of the histogram has changed, the total entropy is the same up to round-off error.



a) Natural Image

b) Histogram

Figure 2.1: The "house" image (of size 240 × 316 and 256 gray levels) and its histogram.



a) Synthetic Image

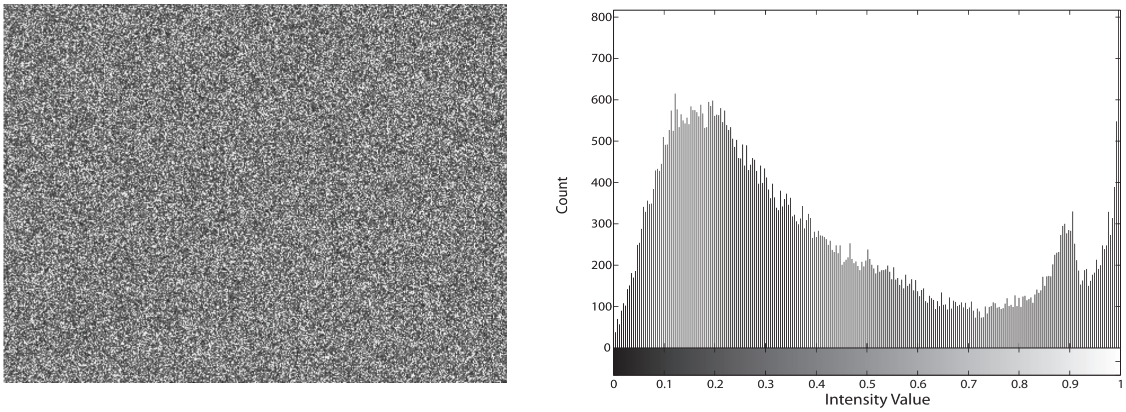
b) Bijective Intensity Mapping

c) Histogram of Synthetic Image

Figure 2.2: A synthetic image created by applying the intensity mapping to the "house" image in previous figure.

Figure 2.3 shows another image created by randomly shuffling the pixel locations of the "house" image. Note that the histogram of both images are the same. Thus, the entropy values computed using (2.5) are the same. However, it is clear that the "house" image contains more structure (i.e., less "uncertainty") and treating each pixel intensity as an independent sample is an oversimplification. In general, it is known that natural images[1] can be successfully modeled as Markov random fields, see e.g. [46]. In this model, pixel intensity values depend on neighboring pixels. In simpler terms, in a natural image the value of a pixel is likely to be close to the value of some of its neighbors. As discussed earlier, this dependency reduces the total entropy of an image, rendering (2.5) an upper bound on the actual entropy.

[1] including medical, scientific, computer-generated images



a) Shuffled Image

b) Histogram

Figure 2.3: A synthetic image created by randomly shuffling the pixel locations in the "house" image and its histogram.

2.2 Other Information-theoretic Measures

While entropy is the basic concept we're going to build our approaches on, it is not the only information-theoretic measure we will be using. We are also interested in relating several random variables and information theory contains many different measures for this purpose. These include conditional entropy, Kullback-Leibler divergence, mutual information and Rényi entropy. In this section, we provide brief descriptions of these measures. In the following X and Y are two discrete random variables with marginal distributions p_X and p_Y and a joint distribution p_{XY}.

2.2.1 Conditional Entropy

Assuming we know the outcome of a random event, conditional entropy is a measure of "new" information gained by observing another event. The formal definition is:

$$H(Y|X) = -\sum_{x,y} p_{XY}(x,y) \log p_{Y|X}(y|x), \tag{2.6}$$

where $p_{Y|X}(y,x) = p_{XY}(x,y)/p_X(x)$ is the conditional distribution. It is easy to see that $H(Y|X) = H(X,Y) - H(X)$.

2.2.2 K-L Divergence

K-L Divergence is a natural distance measure from a distribution $\mathbf{p}$ to another distribution $\mathbf{q}$ and is defined as:

$$D_{KL}(\mathbf{p}||\mathbf{q}) = \sum_i p_i \log \frac{p_i}{q_i}. \tag{2.7}$$

This definition can straightforwardly be extended to the continuous case. Note that $D_{KL}(\mathbf{p}||\mathbf{q}) \geq 0$ and is zero if and only if $\mathbf{p} = \mathbf{q}$. It is not symmetric, i.e., in general $D_{KL}(\mathbf{p}||\mathbf{q}) \neq D_{KL}(\mathbf{q}||\mathbf{p})$, and does not satisfy the triangle inequality. Thus it is not a proper distance metric.

2.2.3 Mutual Information

Mutual Information is a measure of the statistical dependency between two (or more) random variables. It can also be viewed as a measure of the "shared" (common, mutual) information between information sources. A formal definition is:

$$I(X,Y) = \sum_{x,y} p_{X,Y}(x,y) \log \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}. \tag{2.8}$$

Notice that mutual information is equal to the K-L divergence from the joint distribution p_{XY} to the product of the marginals $p_X p_Y$, i.e., the joint distribution when X and Y are independent. Thus $I(X,Y) \geq 0$ and achieves zero if and only if X and Y are independent. Some other important properties of mutual information are:

$$\begin{aligned} I(X,Y) &= I(Y,X) \text{ and I(X, X) = H(X)} \\ I(X,Y) &= H(X) + H(Y) - H(X,Y) = H(Y) - H(Y|X) \qquad (2.9) \\ 0 &\leq I(X,Y) \leq \min\{H(X), H(Y)\} \leq H(X,Y). \end{aligned}$$

2.2.4 Rényi Entropy

This is a generalization of the Shannon entropy and for $\alpha \geq 0$ is defined as [73]:

$$\begin{aligned} H_\alpha(X) &\triangleq \frac{1}{1-\alpha}\log(\sum_x p_X(x)^\alpha) \\ &= \frac{1}{1-\alpha}\log \mathbb{E}_{p_X}(p_X^{\alpha-1}), \end{aligned} \tag{2.10}$$

where $\mathbb{E}_p$ denotes expectation over p. Equation (2.10) can be used to generalize the definition to the continuous case. Note that the limit of H_α as α goes to 1 is Shannon's entropy, H (for a proof, see Appendix A).

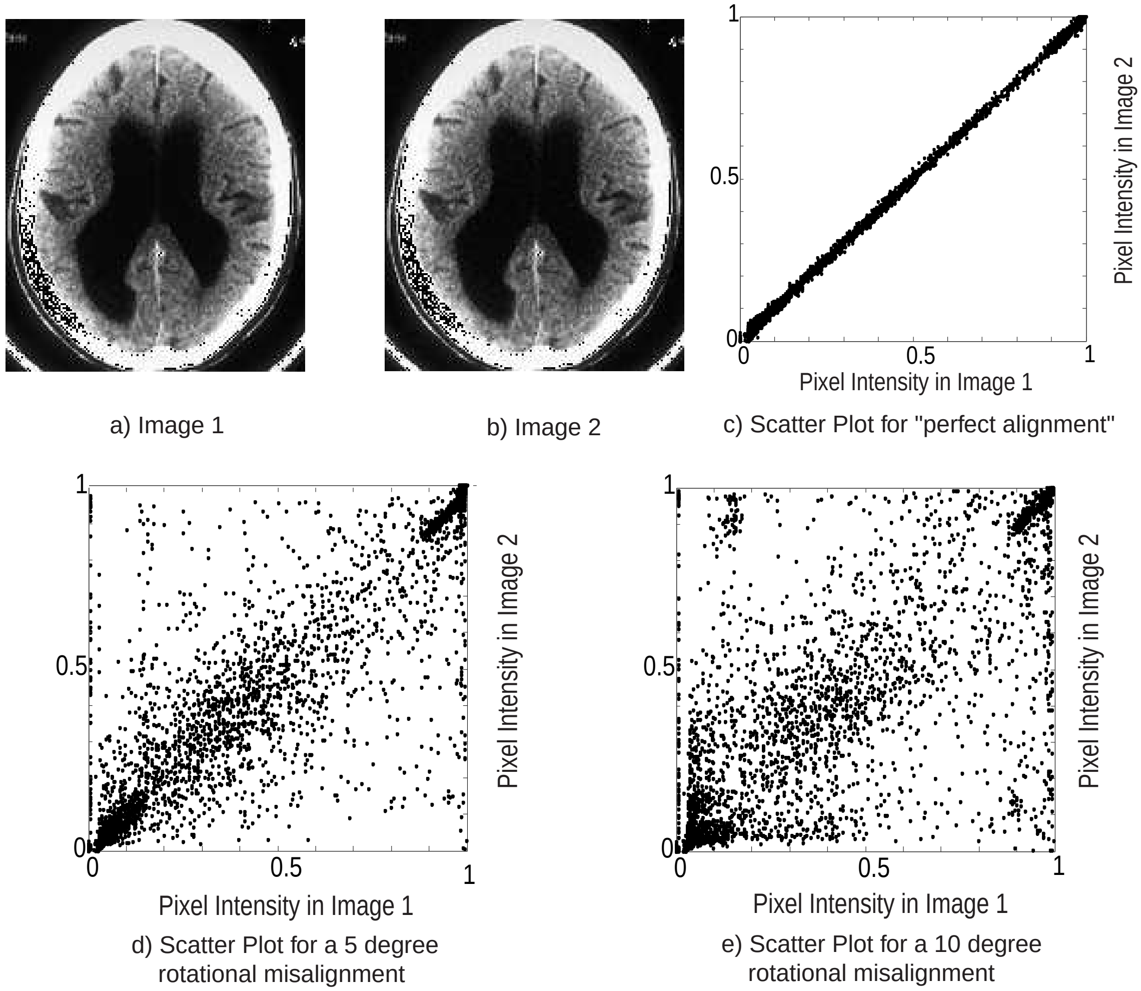
2.3 Entropy as an Alignment Measure

The investigation of information-theoretic measures for image registration started in the 1990's with Woods et al.'s seminal paper [98]. This also marks the beginning of the exploration of fast and reliable *automatic* multi-modal registration methods. The common trait of these approaches is that they rely on the whole image, particularly pixel/voxel intensity values, when determining the quality of alignment. This is contrary to landmark-based approaches that require the definition and computation of specific landmarks. These algorithms are constrained by the quality and speed of the landmark detection step.

The basic idea that motivates the employment of information-theoretic measures for quantifying the quality of alignment is simple: *Corresponding features extracted from the images should become statistically more dependent with better alignment.* This observation is illustrated with the toy example[2] shown in Figure 2.4, where the scatter-plots display pixel intensity value pairs (from both images). Notice that, since both images are the same (up to some noise), at perfect alignment pixel samples

[2] Image obtained from $http://www.med.harvard.edu/JPNM/TF93\_94/Oct12/CT.GIF$

cluster around the $x = y$ line. At bad alignment, however, the samples are scattered, i.e., the joint histogram is more dispersed.



a) Image 1

b) Image 2

c) Scatter Plot for "perfect alignment"

d) Scatter Plot for a 5 degree rotational misalignment

e) Scatter Plot for a 10 degree rotational misalignment

Figure 2.4: Image 1 is a brain CT scan. Image 2 is a synthetic image obtained by corrupting Image 1 by additive i.i.d. Gaussian noise.

In an attempt to quantify the dispersion of the joint histogram both Collignon et al. [13] and Studholme et al. [89] proposed to employ the entropy of the joint histogram for determining alignment quality. These studies were mainly based on the empirical observation that the joint distribution tends to be sharper with well-defined peaks at good alignment, which yields a small entropy. Experiments indicated that the approach was promising, yet no rigorous theoretical derivation was provided.

The papers of Collignon et al.'s [50] and Viola and Wells [97] formalized these ideas and motivated mutual information as an alignment measure. Along with the joint entropy term, the mutual information (2.9) formula includes marginal entropy terms. As argued by many authors, e.g. in [96], this makes mutual information a more suitable alignment measure where there's limited scene overlap between images. In the following years, the basic idea of employing entropy-based measures for various multi-modal image registration applications yielded many successful algorithms. For a detailed overview, see [68]. In the following sections, we provide a theoretical derivation of entropy-based alignment measures. Similar discussions can be found in many other studies, including [96, 75, 103].

2.3.1 Maximum Likelihood and Entropy

In this section, we will provide a maximum likelihood based derivation of entropic alignment measures. Let $U^* : \mathbb{R}^d \mapsto \mathcal{R}$ and $V^* : \mathbb{R}^d \mapsto \mathcal{R}$ be *spatially aligned* scalar images of the same scene, where d is a positive integer and $\mathcal{R}$ is a finite subset of $\mathbb{R}^+$. For example, U^* can be an ultrasound image of one subject's brain and V^* can be an magnetic resonance (MR) image of the same brain. Since, these two images represent the same physical reality, we will model their relationship as:

$$U^* = f(V^*) + E^*, \tag{2.11}$$

where f is a fixed (but typically unknown) mapping from one modality to another and $E : \Omega \mapsto \mathcal{R}$ is some random noise that captures the uncontrollable variables of the imaging process, e.g. magnetic field inhomogeneities in MR. In other words, we are assuming that given f and V^*, we can come up with a reasonable guess for U^*. This is similar to the underlying idea of synthetic data generators, such as Brainweb [12].

In general, any two images, $U : \Omega \mapsto \mathcal{R}$ and $V : \Omega \mapsto \mathcal{R}$, of the same scene are not in spatial alignment. Thus, similar to (2.11), we can write their relationship as:

$$U = f(V \circ \Phi) + E_\Phi, \tag{2.12}$$

where $\Phi : \mathbb{R}^d \mapsto \mathbb{R}^d$ is a geometric transformation. Notice that in (2.12), E_Φ is not only the imaging noise, but also includes the misalignment error (and thus depends on Φ). So, for any Φ, there is a unique E_Φ that satisfies (2.12). If we fix $U = U^*$, then the goal of a registration algorithm is to find $\Phi = \Phi^*$ such that $V \circ \Phi^* = V^*$.

Note that in (2.12) we are not restricting f to be a mapping on intensity values only. In practice, however, to make the problem tractable we put further constraints on this model. For example, in a mono-modal application we may fix f to be an identity mapping and E to be *i.i.d.* Gaussian noise. However, the identity mapping assumption is not suitable for multi-modal applications. A common approach is to treat each pixel independently, and assume f is a mapping on intensity values. Yet, as discussed in [74], in most applications, for example in an Ultrasound-Magnetic Resonance registration application, the assumption of f being a mapping on intensity values is too restrictive. ultrasound images mainly contain "gradient" information, because ultrasound echoes proportional to the difference between acoustical impedances of neighboring tissues. MR images, on the other hand, visualize regions based on hydrogen density. Thus, for the ultrasound-MR application, it is more suitable to assume that f is a mapping on the intensity *and* gradient values of each pixel.

In the following, we will employ a maximum likelihood approach to determine Φ^*. In this framework, we model U, V, Φ and E as random variables and make the following assumptions:

1. *$P(\Phi)$ is a uniform distribution on all feasible transformations $\mathcal{W}$.* For example, if Φ is modeled as a rigid-body transformation, all admissible rigid-body

transformations are considered to be equally likely. This is a strong assumption when working with a rich class of transformations, e.g. nonlinear deformations. Thus, the alignment measure needs to be *regulated*. A detailed discussion of this issue is presented in Chapter 6.

2. Since we fixed $U^* = U$, $P(U|\Phi) = P(U)$, i.e., the image U does not depend on the alignment of the two images. This is a reasonable assumption, since the orientation of U depends only on the orientation of its sensor with respect to the scene.

3. *E_Φ is i.i.d. within a finite region of interest $\Omega \subset \mathbb{R}^d$.* Note that for modeling physical noise due to the imaging process, this may not be a too restrictive assumption. In other words, it is reasonable to assume $E^* = E_{\Phi^*}$ to be i.i.d. However, in (2.12) for an arbitrary Φ; E_Φ also captures the error due to misalignment. Assuming that a *misalignment* error at a pixel is independent of its neighbors is a strong supposition, that will bias the likelihood function downwards.

4. *$f : \mathcal{R} \mapsto \mathcal{R}$ is a bivariate function of intensity values*, i.e.

$$U(\mathbf{x}) = f(V(\Phi(\mathbf{x}))) + E(\mathbf{x}) \tag{2.13}$$

   Recall that $\mathcal{R}$ is the range of image intensity values and is a finite subset of $\mathbb{R}^+$. As discussed above, (2.13) in general is not a good model and the validity of the assumption depends on the application. Yet, in some cases, e.g. [74], one can alleviate the issue by incorporating other features, e.g. the image gradient. We will briefly discuss this in Section 2.4.4.

Based on the above and a generalized maximum likelihood framework, we formulate the registration problem as:

$$\begin{aligned}
\hat{\Phi} &= \arg\max_{\Phi\in\mathcal{W}} \max_{f} P(U,V|\Phi,f) && (2.14)\\
&= \arg\max_{\Phi\in\mathcal{W}} \max_{f} P(V|U,\Phi,f) && (2.15)\\
&= \arg\max_{\Phi\in\mathcal{W}} \max_{f} \sum_{\mathbf{x}\in\Omega} \log P(f(V(\Phi(\mathbf{x})))|U(\mathbf{x})) && (2.16)\\
&\approx \arg\max_{\Phi\in\mathcal{W}} \max_{f} -H(f(V(\Phi(\mathbf{x})))|U(\mathbf{x})) && (2.17)\\
&= \operatorname*{argmin}_{\Phi\in\mathcal{W}} H(V(\Phi(\mathbf{x}))|U(\mathbf{x})), && (2.18)
\end{aligned}$$

where $H(V(\Phi(\mathbf{x}))|U(\mathbf{x}))$ is the conditional entropy of $P(V(\Phi(\mathbf{x}))|U(\mathbf{x}))$. Moving from (2.14) to (2.15) relies on assumption 2. The (2.15-2.16) step employs assumptions 3 and 4; and (2.17) is the asymptotic equipartition property. The approximation relies on $|\Omega|$, the cardinality of the region of interest, to be sufficiently large. The (2.17-2.18) step uses the (2.2) property of entropy. Notice that, in (2.18), the conditional entropy $H(V(\Phi(\mathbf{x}))|U(\mathbf{x}))$ is a misalignment measure (or, equivalently $-H(V(\Phi(\mathbf{x}))|U(\mathbf{x}))$ is an alignment measure).

2.3.2 Fano's Inequality

In this section, we formulate registration as a hypothesis testing problem and provide motivation for entropy-based alignment measures. The analysis is based on the aforementioned model (2.12) and framework presented by Butz et al. in [7]. Note that, we do not rely on any assumptions other than the ones explicitly stated in this section.

Now, we will use a graphical model to explain the imaging process and relate the two images and the scene. Figure 2.5 shows a directed acyclic graph, where blocks represent random variables, arrows indicate (statistical) dependency between the connected r.v.'s. Note that, this graph can also be viewed as a generative process or a first order Markov chain.

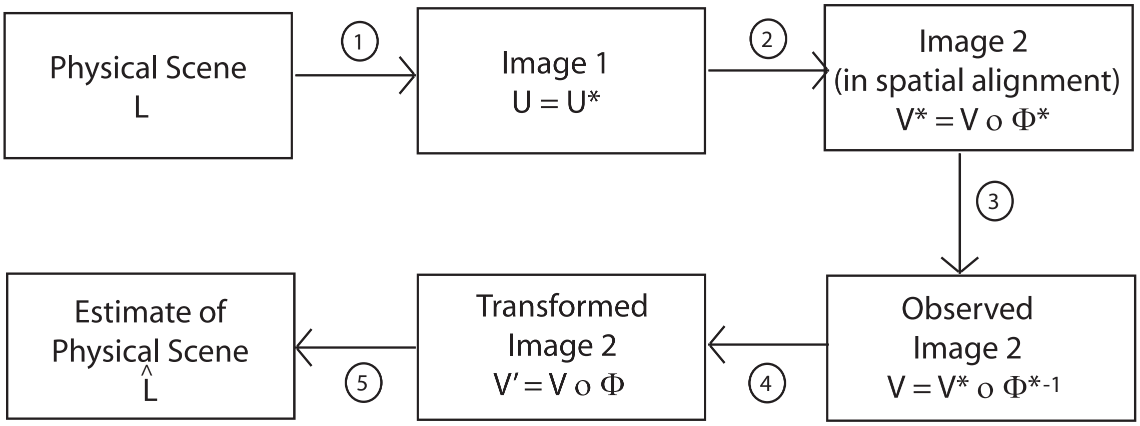


Figure 2.5: A graphical model for relating the two images and the scene.

Let L be a discrete random variable that can take N_L distinct values and represents a physical reality captured by both images U and V. In a medical application, for example, this can represent a tissue type at a fixed location $\mathbf{x}$. In the model (1.1), let's assume the noise is independent of the physical context, i.e., $P(E^*|L) = P(E^*)$. Thus, $P(V^*|U^*, L) = P(V^*|U^*)$, which is the second arrow in Figure 2.5. By Fano's inequality [28] and the data processing inequality [16], we have:

$$\begin{aligned} P(\hat{L} \neq L) &\geq \frac{H(L) - I(L, V') - 1}{\log N_L} && (2.19) \\ &\geq \frac{H(L) - I(U, V') + 1}{\log N_L} && (2.20) \\ &= \frac{H(L) - H(U) + H(U|V') + 1}{\log N_L} && (2.21) \end{aligned}$$

We can consider the probability of error, i.e., $P(\hat{L} \neq L)$ to be a measure of alignment quality. As discussed in [7], based on inequalities (2.20) and (2.21) we can motivate various entropy-based alignment measures. Note that in the given form, the conditional entropy of the images $H(U|V')$ could be an alignment measure, since in the last inequality (2.21), this is the only term that depends on Φ. Given two images, and nothing else, it is, however, impossible to compute $H(U|V)$ without any further assumptions. As before, in (2.12), we can assume f to be a mapping on intensity values and E to be i.i.d noise. Then the conditional entropy $H(U(\mathbf{x})|V(\Phi(\mathbf{x})))$ and mutual information $I(V(\Phi(\mathbf{x})), U(\mathbf{x}))$ can be motivated as alignment measures. Due to the relative abundance of pixel intensity values, these entropy measures are easier to estimate.

Using a similar analysis and a generalized version of Fano's inequality [26, 25], one can motivate Rényi entropy as an alignment measure. This idea will be further developed in Chapter 4.

2.4 Discussion

In previous sections, we presented ways of motivating different entropy measures as alignment measures. The main strength of this approach is the relatively weak assumptions about the inter-image relationship, i.e., the f mapping in the model (2.12). This has led to the success of automatic multi-modal registration algorithms that use information-theoretic alignment measures [68]. However, it is important to note that these algorithms, by no means, provide a universal solution to image registration. They have two important drawbacks:

1. Entropy-based alignment measures are typically computationally more expensive than simpler measures.

2. In most implementations, entropy measures are computed based on the image (and joint) intensity histograms. As discussed in many papers, e.g. [68, 81], this neglect of "spatial information" may affect alignment accuracy.

At this point, we would like to identify some technical details that lead to the drawbacks listed above and briefly discuss practical issues such as speed, accuracy and robustness against noise and initialization. We proceed by dividing the discussion into several categories.

2.4.1 Entropy Estimation

Most of today's information-theoretic registration algorithms estimate the entropic measures, e.g. $H(V(\Phi(\mathbf{x}))|U(\mathbf{x}))$, by modeling sample values, e.g. pixel intensity values, as i.i.d. realizations of a random variable. As discussed in section 2.1.1, it should be noted that this assumption ignores the structure in natural images and biases the entropy estimate.

There are two nonparametric entropy estimation techniques employed for image registration. The popular "plug-in" estimator [3] , which is based on inserting an estimate of the distribution in the entropy expression and has been commonly employed for estimating Shannon's entropic measures [68]. A second, less-known estimation technique, called entropic graphs [40], is based on computing minimal graphs on a set of samples. A monotonic function of the total edge length of these graphs provides a direct estimate of the underlying entropy. This technique has recently been employed for image registration [49, 81]. The employed entropy estimator influences the speed and accuracy of a registration algorithm. To date, there has not been a comprehensive study that has discussed the advantages and disadvantages of the two aforementioned techniques. In succeeding chapters, we attempt to provide a comparison of the two entropy estimators.

2.4.2 Optimization

Registration is an optimization problem where the objective function is the alignment measure and the variables are the transformation parameters. The fact that entropy-based alignment measures tend to have highly non-convex profiles, makes the optimization difficult. Moreover, time may be a constraint of the application and computation of alignment measures can be expensive.

In such cases, a multi-resolution strategy can be used, e.g. [67, 91]. This increases the convergence speed of the optimization algorithm and prevents the algorithm from getting trapped in local extrema. The basic idea is to first solve the problem at a coarse resolution and then use this solution as an initialization with higher resolution representations of the images. An important prerequisite of this strategy is that the optimization scheme should benefit from it. In other words, a good initialization should help speed up the algorithm.

In [13], Collignon et al. use Powell's method for optimization. This method doesn't take full advantage of the multi-resolution strategy since it searches all directions in its direction set at least once, regardless of how close the initial guess is. Hill climbing (gradient descent) methods, on the other hand, e.g. [97], enjoy a dramatic speed-up with good initialization.

2.4.3 Interpolation and Overlap Area

Until this point, we assumed that images are defined on $\mathbb{R}^d$. Digital images, however, have limited spatial resolution, i.e., are defined on a bounded, discrete grid and the method used for interpolating at non-grid locations has a crucial effect on registration accuracy.

There are several well-known techniques to interpolate image intensity values at non-grid locations within the image boundaries. These include bi-linear, bi-cubic, partial-volume and spline-based interpolation methods. [66] contains a detailed analysis of these methods and concludes that the interpolation method determines the sub-pixel accuracy of the algorithm.

There are various ways of handling regions outside image boundaries. A common approach only considers the overlap area between the two images when evaluating an alignment measure. In this implementation, the marginal entropies of both images vary with alignment. Moreover, if a deterministic sampling routine is used, e.g. when using all the pixels, the sample set size is proportional to the overlap area. Thus, entropy-based measures, e.g. conditional entropy of intensity values, can be made arbitrarily small by shrinking the overlap area. One method that addresses this issue normalizes the alignment measure with the sum of the marginal entropies, e.g. as in *normalized mutual information* [88].

In another implementation, a *region of interest*, Ω a finite subset of $\mathbb{R}^d$, is fixed. This is typically the grid of the *fixed image*, U in (2.12). There are two advantages of this approach:

- The marginal entropy of the fixed image is constant with respect to the transformation Φ, which reduces computational complexity.
- The number of samples can be fixed.

A crucial question is then what intensity values to assign in regions outside the boundaries of the *floating image*, V, because this has the potential of introducing superfluous information. One method uses a constant value, usually called constant padding. However, with this we have the freedom to reduce the marginal entropy to zero by transforming the whole image to outside of Ω. Moreover, in practice, we typically want to ensure that near-boundary sample values have a zero gradient along the boundary normal, i.e., we impose a Neumann boundary condition. A simple way of achieving this is to use a nearest neighbor interpolator for out-of-boundary values.

2.4.4 Spatial Information

A drawback of most of today's implementations of entropy-based alignment measures is that they only use pixel intensity values and can typically be computed from the image (and joint) histograms. This is commonly referred to as the *neglect of spatial information*, since pixel locations and the relationship between neighboring pixels don't appear in the alignment measure formula. As an *ad hoc* solution in [66], Pluim et al. propose to combine mutual information with an image gradient-based term that favors similar orientation of edges in both images. Alternatively, Rueckert et al. suggest to use two-dimensional samples (intensity + neighboring intensity) when computing mutual information [78]. Note that, this results in a four-dimensional joint histogram. These studies report that incorporation of neighboring pixel information can increase the robustness of the algorithm, especially against imaging artifacts, such as RF inhomogeneity in MRI, shading, etc.

We observe that the neglect of spatial information is mainly due to two factors:

1. The independence assumption on the samples. This introduces two types of bias in the estimate of the joint entropy (or mutual information) of the two images: one due to the inherent structure of the images. This bias, however, can be assumed to be independent of the image pair alignment and thus is not of our concern. The other source of bias is due to the cross-image dependency of samples, e.g., the intensity of pixel $\mathbf{x}$ in the first image is, in reality, dependent on the intensity of a neighboring pixel in the second image. Moreover, this dependency increases with better alignment.

2. The common supposition that the inter-image mapping f in (2.12) is a function of pixel intensity values only (see assumption 4 in Section 2.3.1). This factor can be alleviated by incorporating knowledge about the imaging modalities and their relationship, as described above in the US-MR example. Note that, a better model can also help with relaxing the independence assumption mentioned in the previous point. For example, we can assume that f is a mapping on 3×3 blocks, which leads us to treat these blocks, instead of pixels, as independent realizations.

However, in general, it is not clear how to relax (2.12) to address the above issues. Thus, we strongly believe that using knowledge about the application domain is critical in improving the performance of an entropy-based registration algorithm and general frameworks will always be constrained by the assumptions. Note that, this point is considered to be out of the scope of this thesis and will not be further discussed.

Chapter 3

Entropy Estimation for Image Registration

In this chapter, we take a short détour from image registration and give a brief overview of different entropy estimation techniques. Until this point, we have generally considered discrete random variables. This was mainly to keep the discussion clear and intuitive. Moreover, when motivating entropy-based alignment measures, we were not worried about issues such as differentiability and interpolated values. For practical purposes, however, we usually model samples extracted from the images as continuous random variables. With common interpolation methods (e.g. bilinear) and most geometric transformations (e.g. rotation), image samples (e.g. pixel intensity values) do, in fact, take on values from a continuous range. In the following section, we (re)define information entropy for a continuous random variable and discuss its relation with the discrete case.

3.1 Entropy of a Continuous Random Variable

The entropy of a continuous random variable is usually called *differential entropy*, and as in the discrete case we denote it by[1] $H(\cdot)$. Let X be a continuous random variable with $p_X(\cdot)$ as its probability density. Shannon's definition [85] of H is:

$$H(X) = H(p_X) \triangleq -\int_{S_X} p_X(x) \log p_X(x) dx, \tag{3.1}$$

where S_X is the support set of p_X. At this point, it is important to note that the maximum likelihood derivation (Section 2.3.1) of entropic alignment measures applies to the continuous case through the extension of the asymptotic equipartition property [16]:

Let $x_1, \ldots, x_n$ be i.i.d samples from p_X, then the log-likelihood function is:

$$-\frac{1}{n} \log p(x_1, \ldots, x_n) \approx H(X).$$

In other words, for a *typical* sequence of samples, the log-likelihood is approximately equal to the entropy of the sequence.

Similar to (3.1), we can define Rényi entropy for the continuous case:

$$H_\alpha(X) = H_\alpha(p_X) \triangleq \frac{1}{1-\alpha} \log \int_{S_X} p_X(x)^\alpha dx, \tag{3.2}$$

where $\alpha > 0$. Once again, it is easy to show that:

$$\lim_{\alpha \to 1} H_\alpha(p_X) = H(p_X).$$

Thus, Rényi entropy (3.2) can be viewed as a generalization of (3.1) (see Appendix A). Note that, in general $H_\alpha(X)$ is not necessarily nonnegative. In fact, for any p.d.f that

[1] Note that, commonly, a small h is used to distinguish differential entropy from the regular discrete entropy. In this thesis, however, we will rely on the context to make that distinction.

contains an impulse, it will achieve $-\infty$. Therefore, many information theorists do not accept differential entropy as a "real" measure of information. Yet, there is a concrete relationship (through quantization) between the discrete and continuous entropies. Here, we will derive this for Rényi entropy. A similar argument for Shannon's entropy can be found in [16].

Let X be a one-dimensional random variable with a sufficiently smooth density p_X. Suppose we divide the support of X into bins of length Δ. By the mean value theorem, there exists a $x_i \in [i\Delta, (i+1)\Delta)$ such that:

$$p_X(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} p_X(x)dx.$$

Define the discrete random variable $X^\Delta = x_i$, if $X \in [i\Delta, (i+1)\Delta)$ with a p.m.f. $p_i = p_X(x_i)\Delta$. Then:

$$\begin{aligned} H_\alpha(X^\Delta) &= \frac{1}{1-\alpha}\log\sum_i p_i^\alpha \\ &= \frac{1}{1-\alpha}\log\sum_i [p_X(x_i)\Delta]^\alpha \\ &= \frac{1}{1-\alpha}\log\sum_i \Delta p_X(x_i)^\alpha - \log\Delta. \end{aligned}$$

If $p(x)^\alpha$ is Riemann integrable, then the first term approaches the Rényi entropy of X, i.e.,

$$H_\alpha(X^\Delta) + \log\Delta \rightarrow H_\alpha(X),$$

as $\Delta \rightarrow 0$. In other words, the Rényi entropy of an n bit quantization of a continuous random variable X is approximately equal to $H_\alpha(X) + n$.

Next, we continue with the problem of estimating (3.1) and (3.2) given a finite set of samples.

3.2 Performance Criteria for Entropy Estimators

Although this thesis deals with the problem of image registration, it is important to note that different entropy-based measures are useful in various other contexts, e.g. [23, 22, 62, 27]. Recall that entropy is defined as a functional of the underlying probability density. In practice, however, algorithms have access to a finite number of samples from that density. Thus, the goal is to approximate, i.e., estimate, the underlying entropy based on the finite sample set.

For a set of i.i.d samples $\{x_1, \ldots, x_n\}$ of the random variable X, let $\hat{H}_n$ denote an estimate of $H(X)$. Then, there are three main types of criteria [3] that summarize the asymptotic behavior of $\hat{H}$:

- Strong consistency: $\lim_{n\to\infty} \hat{H}_n = H(X)$ almost surely.

- Mean square consistency: $\lim_{n\to\infty} \mathbb{E}_X[\hat{H}_n - H(X)]^2 = 0$, where $\mathbb{E}_X$ denotes expectation.

- Weak consistency: $\lim_{n\to\infty} \hat{H}_n = H(X)$ in probability.

In general, we may also be interested in convergence rates, for example upper bounds on $|\mathbb{E}_X \hat{H}_n - H(X)|$ in terms of the number of samples, e.g. [39]. It is known, however, that in a sufficiently rich class of densities, a pdf can always be found such that the entropy estimate converges at an arbitrarily slow rate. For example, no universal convergence rates exist for densities with an infinite support [1]. Thus, in the literature convergence rates are reported for different types and/or restricted classes of distributions [3].

For the purposes of this thesis, we will not be concerned with convergence rates. Rather, we are more interested in how an estimator "treats" data and the best estimator is the one that achieves its optimum at correct alignment and is easy to optimize. The following sections provide an overview of different entropy estimation techniques that fall into two groups: nonparametric and parametric estimators. The latter group relies on the assumption that the underlying p.d.f. belongs to a parameterized family, while the former one does not. Typically, in an image registration problem there is no obvious choice of a parametric density family.

3.3 Nonparametric Entropy Estimation

The only accurate description we can use for nonparametric estimators is that they do *not* model the underlying p.d.f. as a finitely parameterized distribution. In this sense, they are more flexible than parametric methods and are popularly used in settings where we have limited prior knowledge about the "shape" of the underlying distribution. For example, in an image registration application, we usually assume that (at least initially) the images can be arbitrarily badly misaligned. As discussed in the previous chapter, this will cause the samples to be arbitrarily scattered in the scatter plot, e.g. see Figure 2.4. Recall that, we are assuming that these samples are drawn from an underlying (joint) distribution. Generally, there is no natural choice of a parametric density family that this distribution would belong to.

In the following, we assume that $X \in R^d$ is a random variable with the p.d.f $p_X(\mathbf{x})$, $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ is a set of independent samples of X and $|.|$ denotes set cardinality. Here, we include estimates of Rényi entropy $H_\alpha(X)$, which generally can be used for estimating Shannon's entropy by analyzing the limit as α goes to 1. An equivalent measure is the α-information potential [69]:

$$\Lambda_\alpha(X) = \mathbb{E}_X(p_X^{\alpha-1}). \tag{3.3}$$

Without loss of generality, we will consider the problem of estimating (3.3), since $H_\alpha = \frac{1}{1-\alpha} \log \Lambda_\alpha$.

3.3.1 Plug-in Estimators

A plug-in estimator requires a density estimate and the computation of the expectation. We employ a density estimate based on a Parzen-window estimator [21]. This corresponds to using a "blurred" histogram as an estimate of the underlying p.d.f. For $K(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}$, a continuous density, the Parzen estimate of p_X is:

$$\hat{p}_P(\mathbf{x}; \mathcal{X}) = \frac{1}{N} \sum_{i=1}^{N} K(\mathbf{x} - \mathbf{x}_i), \forall \mathbf{x} \in \mathbb{R}^d. \tag{3.4}$$

Note that additional conditions on K determine the convergence rate of this estimate and in practice most kernel functions are symmetric, i.e., $K(\mathbf{x}) = K(-\mathbf{x})$.

Using the Parzen window density estimation approach (3.4), there are at least two ways to approximate the expectation in the entropy expression (3.3). The first approximates the expectation using a sample mean and yields the estimate:

$$\begin{aligned} \hat{\Lambda}_\alpha^M(\mathcal{X}) &\triangleq \frac{1}{N} \sum_{j=1}^{N} \hat{p}_P^{\alpha-1}(\mathbf{x}_j; \mathcal{X}) \\ &= \frac{1}{N^\alpha} \sum_{j=1}^{N} \left(\sum_{i=1}^{N} K(\mathbf{x}_j - \mathbf{x}_i) \right)^{\alpha-1} \end{aligned} \tag{3.5}$$

Strong consistency of this type of estimator was shown in [9] for Shannon's entropy. Note that, in (3.5) we are using the whole sample set to compute both the density and expectation. This is usually called a *re-substitution estimate* [3]. An alternative strategy is to divide the sample set into two mutually exclusive subsets; estimate the density on one and compute the sample mean on the other. This technique is called *splitting data* and was employed by Viola et al. in [97].

The second approach, a histogram-based method, attempts to approximate the infinite integral in the expectation of (3.3) using a simple numerical integration method that approximates the density as a constant within each histogram bin [32]. This was applied to image registration in [91]. If we ignore the nonlinearity introduced by "binning" (i.e., quantizing) and impose some conditions on K, the histogram-based estimate $\hat{\Lambda}^H_\alpha(\cdot)$ is an approximation of the sample mean estimate $\hat{\Lambda}^M_\alpha(\cdot)$ (see Appendix B for details):

$$\hat{\Lambda}^H_\alpha(\mathcal{X}) = \hat{\Lambda}^M_\alpha(q(\mathcal{X})),$$

where $q(\cdot)$ is a quantizer.

In the following, we focus our analysis on a sample mean estimate of the α information potential (3.5). If (3.4) is consistent, it is easy to show that the above estimators, $\hat{\Lambda}^M_\alpha(\cdot)$ and $\hat{\Lambda}^H_\alpha(\cdot)$ are consistent estimators of Λ_α [25].

3.3.2 m-Spacings Estimate

This technique was originally developed for one-dimensional samples [3], i.e. $d = 1$ and makes use the order statistics: $\mathbf{x}_{n,1} \leq \mathbf{x}_{n,2} \leq \ldots \leq \mathbf{x}_{n,n}$. Based on the m-order spacings, $\mathbf{x}_{n,i+m} - \mathbf{x}_{n,i}$, one can construct a density estimate:

$$\hat{p}_S(\mathbf{x}; \mathcal{X}) = \frac{m}{n} \frac{1}{\mathbf{x}_{n,im} - \mathbf{x}_{n,(i-1)m}}, \text{ for } \mathbf{x} \in [\mathbf{x}_{n,(i-1)m}, \mathbf{x}_{n,im}). \tag{3.6}$$

As discussed in [3], consistent entropy estimates can be constructed from (3.6). Recently, Miller extended this technique to higher dimensions using Voronoi regions and Delaunay triangulations [54]. The consistency of this estimate is yet to be shown.

3.3.3 Entropic Spanning Graphs

This technique estimates entropic measures by computing minimal graphs (e.g. a Euclidean minimum spanning tree) on independent samples from the density. In [72], Redmond and Yukich provide a general framework to obtain convergence results for some Euclidean length functionals of specific graphs and this approach has been recently applied to image registration in [49, 81].

Let $G = (E, \mathcal{X})$ be a graph with edge set E and vertex set $\mathcal{X}$. Each graph edge $e = (\mathbf{x}_i, \mathbf{x}_j) \in E$ can be weighted by its Euclidean length $\|e\| = \|\mathbf{x}_i - \mathbf{x}_j\|$. Let $\mathcal{G}_c(\mathcal{X})$ denote a family of graphs with fixed vertex set $\mathcal{X}$ that conforms to a topological constraint c which might be spanning trees, k-neighbor graphs, Hamiltonian cycles, etc (see Appendix for definitions. For a more detailed treatment of entropic graphs see [40, 15, 101, 72]). Normally, c is fixed and understood. Hence, it will not be explicitly indicated. For a $\gamma \in \mathbb{R}$ and graph G, let $W_\gamma(G) \triangleq \sum_{e \in E} \|e\|^\gamma$ denote the *power-weighted graph weight*. For a fixed $\mathcal{G}(\mathcal{X})$, define the *minimum graph weight* (MGW) to be:

$$W^*_\gamma(\mathcal{X}) \triangleq \min_{G \in \mathcal{G}(\mathcal{X})} W_\gamma(G), \tag{3.7}$$

and let

$$G^*(\mathcal{X}) \triangleq \operatorname*{argmin}_{G \in \mathcal{G}(\mathcal{X})} W_\gamma(G) \tag{3.8}$$

denote a *minimal graph*. Note $W_\gamma(G^*(\mathcal{X})) = W^*_\gamma(\mathcal{X})$ and $G^*(\mathcal{X})$ is not necessarily unique.

In the following, we provide a result that allows the MGW (3.7) to be employed for entropy estimation. Let p_X be a Lebesgue density on $[0,1]^d$ and $\gamma = d(1-\alpha)$. Set:

$$\tilde{H}_\alpha(\mathcal{X}) \triangleq \frac{1}{1-\alpha} \log \left(\frac{W^*_{d(1-\alpha)}(\mathcal{X})}{N^\alpha} \right). \tag{3.9}$$

In [41], Hero et al. show that for all $\alpha \in (0,1)$, $\tilde{H}_\alpha(\mathcal{X}) + \frac{\log \beta}{1-\alpha}$ is a strongly consistent estimator of $H_\alpha(X)$ (as the sample size grows to infinity), where β is a (generally unknown) constant that depends on the topological constraint c, the parameter α and the sample dimension d, but *not* on p_X. Correspondingly, a graph-theoretic estimate of the α information potential is:

$$\hat{\Lambda}^G_\alpha(\mathcal{X}) \triangleq \beta \frac{W^*_{d(1-\alpha)}(\mathcal{X})}{N^\alpha}. \tag{3.10}$$

Now, we show a new result that recognizes the entropic graph estimate as a special case of the plug-in method. To show this, let's define $A(G)$ as the adjacency matrix of the graph G, where the non-diagonal (ij)'th entry $A(G)(i,j)$ is the number of edges joining vertex i and vertex j. As discussed in the appendix, for a TSP, MST or nearest neighbor graph there exists a matrix $L(G)$ such that $L(G) + L(G)^T = A(G)$, where L^T denotes transpose of L, and each row of $L(G)$ has at most one non-zero entry that is equal to 1. Using this fact, we can write:

$$\begin{aligned}
W^*_\gamma(\mathcal{X}) &= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \|\mathbf{x}_i - \mathbf{x}_j\|^\gamma A(G^*(\mathcal{X}))(i,j) \\
&= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \|\mathbf{x}_i - \mathbf{x}_j\|^\gamma \{L(G^*(\mathcal{X}))(i,j) + L(G^*(\mathcal{X}))^T(i,j)\} \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N} \|\mathbf{x}_i - \mathbf{x}_j\|^\gamma L(G^*(\mathcal{X}))(i,j) \\
&= \sum_{i=1}^{N}\left[\sum_{j=1}^{N} \|\mathbf{x}_i - \mathbf{x}_j\|^\beta L(G^*(\mathcal{X}))(i,j)\right]^{\gamma/\beta}. \qquad (3.11)
\end{aligned}$$

Inserting (3.11) into (3.10), we can rewrite the entropic graph estimate (for MST's, NN's, or TSP's) of the α information potential as:

$$\hat{\Lambda}^G_\alpha(\mathcal{X}) = \beta\frac{W^*_{d(1-\alpha)}(\mathcal{X})}{N^\alpha} = \frac{1}{N}\sum_{i=1}^{N} \hat{p}_G(\mathbf{x}_i; G^*(\mathcal{X}))^{\alpha-1}, \qquad (3.12)$$

where $\hat{p}_G(\mathbf{x}_i; G^*(\mathcal{X})) \triangleq \frac{\beta'}{N}\sum_{j=1}^{N} \|\mathbf{x}_i - \mathbf{x}_j\|^{-d} L(G^*(\mathcal{X}))(i,j)$ and $\beta' = \beta^{\frac{1}{\alpha-1}}$.

If G is a K-NN graph (i.e., $G = \cup_{k=1}^{K} G_k$, where G_k is the kth nearest neighbor graph) the estimator is:

$$\hat{\Lambda}^{GK}_\alpha(\mathcal{X}) = \frac{\beta}{K}\frac{W^*_{d(1-\alpha)}(\mathcal{X})}{N^\alpha} = \frac{1}{KN}\sum_{k=1}^{K}\sum_{i=1}^{N} \hat{p}_G(\mathbf{x}_i; G^*_k(\mathcal{X}))^{\alpha-1}, \qquad (3.13)$$

where $\hat{p}_G(\mathbf{x}_i; G^*_k(\mathcal{X})) \triangleq \frac{\beta'}{N}\sum_{j=1}^{N} \|\mathbf{x}_i - \mathbf{x}_j\|^{-d} L_k(G^*_k(\mathcal{X}))(i,j)$. Notice that the K-NN estimator is an averaged version of the NN estimator.

Entropic Spanning Graphs as Plug-in Estimators

By deriving the relevant expressions in a common framework, we showed that the entropic graph estimate (3.12) is a special case of the plug-in estimator (3.5). In the plug-in estimator the whole sample set is employed when evaluating the underlying probability density at a sample value. On the other hand, the entropic graph estimator employs only one (or a subset) of the closest neighbors to evaluate the density at a sample value. This density estimate can be viewed as uniformly distributing the sample probability over a ball around each sample. The radius of this ball is equal to the Euclidean distance to the relevant neighbor, $\|\mathbf{x}_i - \mathbf{x}_j\|$, and the volume of the ball is proportional to $\|\mathbf{x}_i - \mathbf{x}_j\|^d$. Thus, one interpretation is that the entropic graph estimator uses a variable width kernel that locally adapts to the data, whereas the plug-in estimator employs a global, constant-width kernel.

3.4 Parametric Entropy Estimation

In some cases, we have a good idea about the form of the underlying distribution. For example, we may know that the samples are drawn from a Gaussian distribution. For many distributions, including the Gaussian, closed form expressions exist for the differential entropy (see Chapter 16 of [16] for a list). Then, inserting maximum likelihood (ML) estimates of the distribution parameters into these expressions yields entropy estimates.

The differential entropy of a multivariate Gaussian is:

$$\frac{1}{2}\log((2\pi e)^d|\Sigma|),$$

where $|\Sigma|$ is the determinant of the covariance matrix and e is the natural number.

Note that the ML estimate of Σ is: $\hat{\Sigma} = \frac{1}{N}\sum_i (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$, where $\bar{\mathbf{x}} = \frac{1}{N}\sum_i \mathbf{x}_i$ is the sample mean.

Similarly, for a set of one dimensional i.i.d. samples $\mathcal{X} = \{x_1, \ldots, x_N\}$ from an exponential distribution:

$$p(x) = \frac{1}{\lambda}e^{-\frac{x}{\lambda}},\ x, \lambda > 0,$$

an estimate of the differential entropy is $1 + \log\hat{\lambda}$, where $\frac{1}{\hat{\lambda}} = \frac{1}{N}\sum_i x_i$.

For a Laplace distribution:

$$p(x) = \frac{1}{2\lambda}e^{-\frac{|x-\theta|}{\lambda}},\ \lambda > 0,$$

an entropy estimate is $1 + \log(2\hat{\lambda})$, where $\hat{\lambda} = \frac{1}{N}\sum_i |x_i - \hat{\theta}|$ and $\hat{\theta} = \text{median}(\mathcal{X})$.

Chapter 4

Rényi Entropy-based Image Registration

In this chapter, we return to image registration and discuss the employment of Rényi entropy as a misalignment measure. In particular, we motivate the use of this measure and develop an efficient graph theoretic algorithm that jointly estimates Rényi entropy and its descent direction structure with respect to a parameterized class of spatial transformations. We then provide both a theoretical and practical comparison of the proposed algorithm with the popular plug-in estimator. We highlight the similarities between their gradients (or descent directions) and discuss the practical implications of the variations on a registration algorithm's performance.

4.1 Rényi Entropy as a Misalignment Measure

Using a similar approach to the analysis reported in [8] and described in Section 2.3.2, we can motivate Rényi entropy as a misalignment measure. This requires a generalized version of Fano's inequality, which was recently derived by Erdogmus and Principe and is reported in [25]. Once again, let U and V be two images of the same scene. U and V are in general not in spatial alignment. Without loss of generality, we build the following Markov Chain:

$$U \rightarrow V \rightarrow \hat{U},$$

where the first link represents the valumetric and geometric variations between the two images (see Chapter 1 for a qualitative description of this idea) and $\hat{U} = \hat{f}(V)$, where $\hat{f}$ is an estimate of f in Equation (2.12). Then, the probability of error, i.e., $P(\hat{U} \neq U)$ can be considered as an indicator of how well the images are spatially aligned.

A generalized version of Fano's inequality provides upper and lower bounds on the error probability. Assuming each pixel value is an i.i.d. sample of a discrete random variable:

$$\begin{aligned} \frac{H_\beta(V(\mathbf{x}), U(\mathbf{x})) - H(V(\mathbf{x})) - 1}{\log(N_q - 1)} &\leq P(\hat{U} \neq U) && (4.1) \\ &\leq \frac{H_\alpha(V(\mathbf{x}), U(\mathbf{x})) - H(V(\mathbf{x}))}{\min_k H(U(\mathbf{x})|\hat{U} \neq U, V(\mathbf{x}) = k)} && (4.2) \\ &\leq \frac{H_\alpha(V(\mathbf{x}), U(\mathbf{x}))}{\min_k H(U(\mathbf{x})|\hat{U} \neq U, V(\mathbf{x}) = k)}, \end{aligned}$$

$\forall \alpha \in (0,1), \forall \beta \geq 1$, and where H_α and H_β are Rényi entropies, H is Shannon's entropy and N_q is the number of possible intensity values for U. In [27], the author indicates that the bounds are tighter for α and β close to 1. Moreover, the denominator in the upper bound is maximum (and thus the bound is the tightest) when the probability on $U(\mathbf{x})$ is uniformly distributed over the *wrong* values.

Inspired by the upper bound in (4.2), in the remainder of this chapter, we investigate the joint Rényi entropy, $H_\alpha(U(\mathbf{x}), V(\mathbf{x}))$, for $\alpha \in (0, 1]$, as a misalignment measure. However, before moving on, it is important to identify some potential problems with this approach. Ignoring the denominator in the upper bound makes the alignment measure sensitive to the overlap area and initial alignment, because the denominator term can be made as small as possible in non-overlapping areas. Consider the following example: The set $A_k = \{\mathbf{x} : \Phi(\mathbf{x}) \in \Omega_V \text{ and } U(\mathbf{x})) = k\}$ is empty for some $k \in \{1, \ldots, N_q\}$. If we assign a constant value to $V(\Phi(\mathbf{x}))$ in out-of-boundary regions, i.e., outside of Ω_V, then the denominator becomes equal to zero. This renders the upper bound redundant and the entropy measure is not a useful alignment measure.

To handle this issue, Studholme et al. propose to *normalize* joint entropy with the sum of the marginal entropies, $H(U)+H(V)$ and compute the measures in the overlap area [88]. The normalization makes the alignment measure invariant to overlap area.

Alternatively, as discussed in Section 2.4.3, we can compute the alignment measure on a fixed region of interest, making the marginal entropy of the fixed image U constant, and use an appropriate interpolator (e.g. nearest neighbor) for out-of-bound values. Note that, this method addresses the limited overlap problem and is suitable for volume-preserving transformations, e.g. rigid body. However, it fails when there's scaling in the transformation. For example, consider the extreme case of blowing up V such that the whole region of interest falls into one pixel/voxel. Then $H_\alpha(U, V \circ \Phi)$ is minimized and is equal to $H_\alpha(U)$. As discussed in various studies, e.g. [96, 102], the marginal entropy terms in mutual information handle this issue. Similarly, based on the numerator in the upper bound of (4.2), one can use:

$$H_\alpha(U(\mathbf{x}), V(\Phi(\mathbf{x}))) - H(V(\Phi(\mathbf{x}))), \tag{4.3}$$

as a misalignment measure with scaling transformations, where H_α is Rényi's entropy, $\alpha \in (0, 1]$ and H is Shannon's entropy.

4.2 Rigid Registration

Here, we consider a 3D to 3D registration problem, i.e., $\Phi : \mathbb{R}^3 \mapsto \mathbb{R}^3$. In a rigid-body registration algorithm, the transformation has six parameters: three for rotation

(α, β, γ) and three for translation (t_x, t_y, t_z). Let $\mathbf{r} = [t_x, t_y, t_z, \alpha, \beta, \gamma]$. The transformation can be expressed as:

$$\Phi(\mathbf{x}; \mathbf{r}) = R \times (\mathbf{x} - \mathbf{c}) + \mathbf{t} + \mathbf{c}, \tag{4.4}$$

where

$$\begin{aligned} R &= \begin{pmatrix} \cos\alpha\cos\gamma + \sin\alpha\sin\beta\sin\gamma & \cos\beta\sin\gamma & -\sin\alpha\cos\gamma + \cos\alpha\sin\beta\sin\gamma \\ -\cos\alpha\sin\gamma + \sin\alpha\sin\beta\cos\gamma & \cos\beta\cos\gamma & \sin\alpha\sin\gamma + \cos\alpha\sin\beta\cos\gamma \\ \sin\alpha\cos\beta & -\sin\beta & \cos\alpha\cos\beta \end{pmatrix}, \\ \mathbf{t} &= \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix}, \end{aligned}$$

and $\mathbf{c} \in \mathbb{R}^3$ is an arbitrary fixed center of rotation. We formulate rigid registration as:

$$\mathbf{r}^* = \underset{\mathbf{r}}{\operatorname{argmin}}\, \hat{\Lambda}_\alpha(U(\mathbf{x}), V(\Phi(\mathbf{x}; \mathbf{r}))), \tag{4.5}$$

where $\hat{\Lambda}_\alpha$ is an estimate of the α information potential (3.3).

4.3 Gradient Descent Optimization

In (4.5), we expressed image registration as an optimization problem. In practice, we can put constraints on the transformation parameters, e.g. they cannot be too large. Today, most fast algorithms that employ information-theoretic alignment measures are variants of gradient descent or ascent, e.g. [91, 68, 51].

Let $\nabla_\mathbf{r}$ denote the gradient w.r.t. $\mathbf{r}$. In the following, let:

$$\mathcal{S}_\mathbf{r} = \{\mathbf{s}_i\} = \{(U(\mathbf{x}), V(\Phi(\mathbf{x}; \mathbf{r}))), \forall \mathbf{x} \in \Omega\}, \tag{4.6}$$

and $N = |\Omega|$. Using the chain rule, the gradient of $\hat{\Lambda}$ can be written in the following form:

$$\nabla_{\mathbf{r}} \hat{\Lambda}_\alpha(\mathcal{S}_{\mathbf{r}}) = \sum_{i=1}^{N} \nabla_{\mathbf{r}} \mathbf{s}_i \dot{\nabla}_{\mathbf{s}_i} \hat{\Lambda}_\alpha(\mathcal{S}_{\mathbf{r}}). \tag{4.7}$$

The second term in the summation is a 2-dimensional gradient vector of the misalignment measure with respect to sample values. The first term is the Jacobian of the sample value with respect to the transformation parameters and depends on the images, the interpolation method and geometric transformation, but not the misalignment measure. Hence, only the second term is of interest when comparing different misalignment measures.

In the following sections, we derive and compare "gradient" expressions for two estimates of Λ_α, namely the entropic spanning graph and the plug-in estimator.

4.3.1 Plug-in Estimator

An advantage of sample mean plug-in estimators is that they are readily differentiated. The gradient of (3.5) can be written as:

$$\nabla_{\mathbf{s}_j} \hat{\Lambda}_\alpha^M(\mathcal{S}) \quad = \quad (\alpha - 1) \sum_{k \neq j} n_M(\mathcal{S}, \alpha, j, k) \mathbf{f}_M(\mathbf{s}_j, \mathbf{s}_k), \tag{4.8}$$

where

$$\begin{aligned} n_M(\mathcal{S}, \alpha, j, k) &\triangleq \frac{1}{N^\alpha} \left[(\sum_{l=1}^{N} K(\mathbf{s}_j - \mathbf{s}_l))^{\alpha-2} + (\sum_{l=1}^{N} K(\mathbf{s}_k - \mathbf{s}_l))^{\alpha-2} \right] \\ &= N^2 (\hat{p}(\mathbf{s}_j)^{\alpha-2} + \hat{p}(\mathbf{s}_k)^{\alpha-2}), \end{aligned} \tag{4.9}$$

and

$$\mathbf{f}_M(\mathbf{s}_j, \mathbf{s}_k) \triangleq \nabla K(\mathbf{s}_j - \mathbf{s}_k). \tag{4.10}$$

As will be seen in Section 4.3.3, (4.8) can be viewed as a sum of pairwise attraction terms $\mathbf{f}_M$ weighted by the network terms n_M.

4.3.2 Entropic Spanning Graphs

The entropic graph estimate (see Section 3.3.3 and [40] for definition) is not always differentiable (see Lemma E.0.4 in Appendix E for details). We illustrate this with the following toy example, where $\mathcal{G}$ is the family of spanning trees.

Example 4.3.1. *Consider the vertex set $\mathcal{V} = \{v_1, v_2, v_3\}$ with edges and parameterized lengths (for $-0.3 \le t \le 0.3$): $\|e_{12}\| = \sqrt{(0.3+t)^2 + 0.36}$, $\|e_{23}\| = \sqrt{(0.3-t)^2 + 0.36}$ and $\|e_{13}\| = 0.6$ (see Figure 4.1). It's easy to show that at $t = 0^-$, the MST consists of e_{12} and e_{13}, whereas at $t = 0^+$, e_{23} and e_{13} belong to the MST. Thus $dW(0^-)/dt = 0.6/e_0$ and $dW(0^+)/dt = -0.6/e_0$, where $e_0 = \sqrt{0.45}$. Since the left and right derivatives are not equal, the derivative of the MST weight does not exist at $t = 0$.*

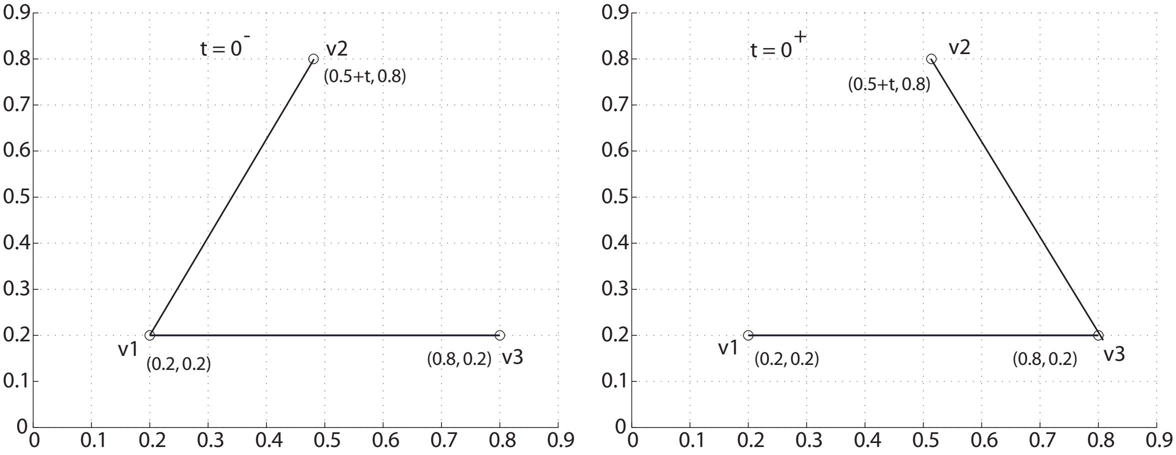


Figure 4.1: A toy example that illustrates the non-differentiability of the EMST weight.

We observe that the non-differentiability of the entropic graph estimate is due to the fact that *the topology of the minimal graph is not constant as the transformation* *parameters are continuously varied.* Originally, this was thought to be an important disadvantage of entropic graph estimators. However, the following result shows that a descent direction of the total weight of any of the minimal graphs, $W_\gamma(G^*)$ is also a descent direction for the MGW, W^*_γ (3.7), and equivalently the misalignment measure. We short-hand $W^*_\gamma(\mathcal{S}_\mathbf{r})$ with $W^*_\gamma(\mathbf{r})$. Note that, $\mathbf{r}$ is a vector of transformation parameters and in general, can be of any size, i.e., $\mathbf{r} \in \mathbb{R}^n$ for $n \in \mathbb{Z}^+$. Recall that for a $3D$ rigid body transformation $n = 6$.

Theorem 4.3.2. *Let $G^*(E^*_\mathbf{r}, \mathcal{S}_\mathbf{r})$ be a minimal graph over $\mathcal{S}_\mathbf{r}$ and $\mathbf{u} \in \mathbb{R}^n$ be a unit vector. Then, if*

$$\sum_{e \in E^*_{\mathbf{r}_0}} \nabla_\mathbf{r} \|e(\mathbf{r}_0)\|^\gamma \cdot \mathbf{u} \tag{4.11}$$

exists and is negative, then $\exists \epsilon > 0$ such that $W^*_\gamma(\mathbf{r}_0 + h\mathbf{u}) \le W^*_\gamma(\mathbf{r}_0)$ for all $0 \le h \le \epsilon$.

Proof: If (4.11) exists and is negative, by vector calculus $\exists \epsilon > 0$ such that:

$$\sum_{e \in E^*_{\mathbf{r}_0}} \|e(\mathbf{r}_0 + h\mathbf{u})\|^\gamma \le \sum_{e \in E^*_{\mathbf{r}_0}} \|e(\mathbf{r}_0)\|^\gamma = W^*_\gamma(\mathbf{r}_0), \tag{4.12}$$

for all $0 \le h \le \epsilon$. By definition, we have:

$$W^*_\gamma(\mathbf{r}_0 + h\mathbf{u}) \le \sum_{e \in E^*_{\mathbf{r}_0}} \|e(\mathbf{r}_0 + h\mathbf{u})\|^\gamma. \tag{4.13}$$

Hence combining (4.12) and (4.13), we get $W^*_\gamma(\mathbf{r}_0 + h\mathbf{u}) \le W^*_\gamma(\mathbf{r}_0)$.□

Choose a minimal graph $G^*(\mathcal{S}_{\mathbf{r}_0})$. Define:

$$\mathbf{d}_\gamma(G^*(\mathcal{S}_{\mathbf{r}_0})) \triangleq -\nabla_\mathbf{r} W_\gamma(G^*(\mathcal{S}_{\mathbf{r}_0})) = -\sum_{e \in E^*_{\mathbf{r}_0}} \nabla_\mathbf{r} \|e(\mathbf{r}_0)\|^\gamma, \tag{4.14}$$

the steepest descent direction for the chosen $W_\gamma(G^*)$. It is easy to see that, when nonzero and finite, $\mathbf{d}_\gamma / \|\mathbf{d}_\gamma\|$ satisfies the condition in (4.11) and therefore is a descent direction for W^*_γ. Note that, if zero length edges exist, i.e., some sample values coincide, and $\gamma < 1$, then (4.14) does not exist and (4.11) is never satisfied. In practice, the direction we choose for this problematic case is:

$$\bar{\mathbf{d}}_\gamma(G^*(\mathcal{S}_{\mathbf{r}_0})) \triangleq - \sum_{e \in E^*_{\mathbf{r}_0}, \|e\| \neq 0} \nabla_{\mathbf{r}} \|e(\mathbf{r}_0)\|^\gamma, \tag{4.15}$$

which is the steepest descent direction for the graph that excludes the zero-length edges, i.e., the minimal graph on unique samples. Note that $\bar{\mathbf{d}}_\gamma = \mathbf{d}_\gamma$, when $\mathbf{d}_\gamma$ exists and is finite.

More complex schemes for finding a descent direction are also possible, e.g. selecting several G^*'s and averaging the corresponding descent directions. However, we focus our analysis on the descent direction obtained from one of the minimal entropic graphs G^*. Correspondingly, for a fixed $G^*(\mathcal{S})$, we define the *pseudo-gradient*, $\mathbf{g}_j(G^*(\mathcal{S}))$, of the entropic graph estimate of the α-information potential, $\hat{\Lambda}^G_\alpha(\mathcal{S})$ (3.10) w.r.t. $\mathbf{s}_j$ as:

$$\mathbf{g}_j(G^*(\mathcal{S})) \triangleq (\alpha - 1) \sum_{\mathbf{s}_k \in \mathcal{S}} n_G(\mathcal{S}, \alpha, j, k) \mathbf{f}_G(\mathbf{s}_j, \mathbf{s}_k); \tag{4.16}$$

where

$$n_G(\mathcal{S}, \alpha, j, k) = \frac{d(a - \alpha)\beta}{2N^\alpha} A(G^*(\mathcal{S}))(j, k) \tag{4.17}$$

is the network weight and

$$\mathbf{f}_G(\mathbf{s}_j, \mathbf{s}_k) = \begin{cases} \|\mathbf{s}_j - \mathbf{s}_k\|^{d-2-d\alpha}(\mathbf{s}_j - \mathbf{s}_k) & \text{if } \|\mathbf{s}_j - \mathbf{s}_k\| > 0, \\ 0 & \text{else,} \end{cases} \tag{4.18}$$

is the sample pair attraction. Recall, $A(G)$ is the adjacency matrix of the graph G. Note, we have put both the gradient of the sample mean plug-in estimator (4.8, 4.9, 4.10) and descent direction (pseudo-gradient)[1] for the entropic graph estimator (4.16, 4.17, 4.18) into a common comparative form involving a pairwise attraction and a

[1] Note that we are assuming a constant topology of the minimal graph

corresponding network weight.

4.3.3 Comparison of the Two Estimators

In this section, we compare the "gradient" expressions for the two entropy estimators. This analysis provides some useful insights on the performance of registration algorithms employing these entropy estimators and their "gradients". In the remainder we make the following common simplifying assumptions:

- The kernel used for the plug-in estimator is a two-dimensional separable Gaussian, $G_\sigma(x, y) = g_\sigma(x) g_\sigma(y)$, where $g_\sigma(\cdot)$ is a zero mean Gaussian with variance σ^2.

- The family of spanning tree graphs is used to compute a minimal entropic graph. A minimum spanning tree (MST) has a 0-1 adjacency matrix, i.e., $A(G^*)(i, j) \in \{0, 1\}$ for all i, j.

- The optimization is an iterative descent scheme and the transformation parameter update can be written in the form: $\mathbf{r}_{m+1} = \mathbf{r}_m + \lambda_m * \sum_j (\sum_k n_{jk} \mathbf{f}_{jk}) \cdot \nabla_{\mathbf{r}} \mathbf{s}_j^m$, where λ_m is a step size, $\mathbf{f}_{jk}$ is the sample pair attraction, n_{jk} is the network weight, $\nabla_{\mathbf{r}} \mathbf{s}_j^m$ is the gradient of the jth sample w.r.t the transformation parameters and $\mathbf{r}_m$ is the value of $\mathbf{r}$ at the mth iteration. n_{jk} and $\mathbf{f}_{jk}$ are summarized (ignoring constants) for the two entropy estimators in Table 4.1. Their product represents the influence of this sample pair interaction on the total gradient, and hence their effect on the gradient.

	Plug-in	Entropic Graph
$\mathbf{f}_{jk}$	$e^{-c\|\mathbf{s}_j - \mathbf{s}_k\|^2} \|\mathbf{s}_j - \mathbf{s}_k\| \mathbf{u}_{jk}$	$\|\mathbf{s}_j - \mathbf{s}_k\|^{1-2\alpha} \mathbf{u}_{jk}$
n_{jk}	$[\hat{p}(s_j)^{\alpha-2} + \hat{p}(s_k)^{\alpha-2}]$	$A(G^*)(j, k) = 1$ or 0

Table 4.1: Comparison of the influence of sample pair ($\mathbf{s}_j$ and $\mathbf{s}_k$) interactions on the update equation. Note $c = 1/2\sigma^2$ and $\mathbf{u}_{jk}$ is the unit vector pointing from $\mathbf{s}_j$ to $\mathbf{s}_k$.

Plug-in Estimator

First, let's consider the sample mean plug-in estimator. The computation time of the estimator and its gradient is $\mathcal{O}(N^2)$, where N is the total number of samples[2]. Figure 4.2a, shows the attraction field to a sample located at the origin, i.e., $\mathbf{f}_M(\mathbf{0},.)$(4.10). With the plug-in estimator, the attraction field does not depend on α, but the network weight does. Also, the attractive force between two samples is zero when they coincide, achieves a maximum value at a close distance σ and becomes negligible when they are far apart.

To analyze the network effect consider a cluster of points, where a cluster can be thought of as a set of points with a relatively small diameter ρ_c. Let $\mathbf{s}_c$ and N_c denote the mean value and number of samples within the cluster, respectively. The total net force[3] generated by this cluster and acting on a sample $\mathbf{s}_j$ is approximately:

$$N_c * N^2 \left[\hat{p}(\mathbf{s}_c)^{\alpha-2} + \hat{p}(\mathbf{s}_j)^{\alpha-2}\right] * e^{-c\|\mathbf{s}_j-\mathbf{s}_c\|^2}\|\mathbf{s}_j - \mathbf{s}_c\|\mathbf{u}_{jc},$$

where $\mathbf{u}_{jc}$ is the unit vector pointing from $\mathbf{s}_j$ to $\mathbf{s}_c$. Assuming all $s \in \mathcal{S}$ are independent samples of a sufficiently smooth density $p(\cdot)$, by the law of large numbers $N_c \propto Np(\mathbf{s}_c)$ and the total net force is approximately proportional to:

$$N^{\alpha+1} * n_M^c(p(\mathbf{s}_c); p(\mathbf{s}_k)) * e^{-c\|\mathbf{s}_j-\mathbf{s}_c\|^2}\|\mathbf{s}_j - \mathbf{s}_c\|\mathbf{u}_{jc}, \tag{4.19}$$

where $n_M^c(p(\mathbf{s}_c); p(\mathbf{s}_k)) = p(\mathbf{s}_c)^{\alpha-1} + p(\mathbf{s}_c)p(\mathbf{s}_j)^{\alpha-2}$ is the total network weight between a cluster and a point. Note that n_M^c is a monotonically increasing function of $p(\mathbf{s}_c)$ when $p(\mathbf{s}_c) > p(\mathbf{s}_j)$, and a monotonically decreasing function of $p(\mathbf{s}_j)$. Thus, we observe that *low probability samples are attracted to high probability, i.e., more crowded,*

[2] Today, most practical entropy-based registration algorithms employ histogram-based fast approximations of the plug-in estimate. Assuming the number of histograms is $\mathcal{O}(N^{1/3})$, this entropy estimate has a computational complexity of $\mathcal{O}(N^{4/3})$

[3] net force equals attraction force times network weight

clusters with a greater force.

Entropic Graphs

The computation time of the entropic graph estimator is $\mathcal{O}(N \log N)$. One advantage of this estimator is that once a minimum entropic graph is computed, the computation of the gradient for any α value is $\mathcal{O}(N)$ and negligible in practice[4]. As inter-sample distance $\|\mathbf{s}_j - \mathbf{s}_k\|$ approaches zero, the sample pair attraction $\mathbf{f}_G(\mathbf{s}_j, \mathbf{s}_k)$ does not converge for $\alpha > 0.5$, but converges to 0 for $\alpha \leq 0.5$. Figure 4.2 shows the attraction field $\mathbf{f}_G(\mathbf{0}, .)$ for the entropic graph estimator with three different α values. When $\alpha > 0.5$, the attraction field achieves arbitrarily large magnitudes around the origin and monotonically decreases at a much slower pace than the plug-in estimator as one moves away from the origin. When $\alpha < 0.5$, however, it is zero at the origin and monotonically increases as one moves away. The network effect, on the other hand, is either 1, if the two samples are connected in the minimal graph; or 0, otherwise. Thus, only a small subset of the sample pair interactions actually influence the gradient.

Samples, Gradients and Image Registration

When digital images are uniformly sampled, coarse structures typically have a large representation, whereas fine detail structures are weakly represented. Thus, with a pair of images, sample clusters typically correspond to partially overlapping coarse image structures. Outliers, i.e., isolated samples that don't belong to a cluster, are usually due to a misaligned region, a point that has no correspondence, or noise. The goal of a registration algorithm can be viewed as "to pull in" outliers toward reliable clusters. Lacking any other useful information, it is natural to trust clusters rather than outliers when driving the registration algorithm.

At bad image alignment, we expect samples from fine detail structures to have

[4] This is true for MST and kNN, not TSP

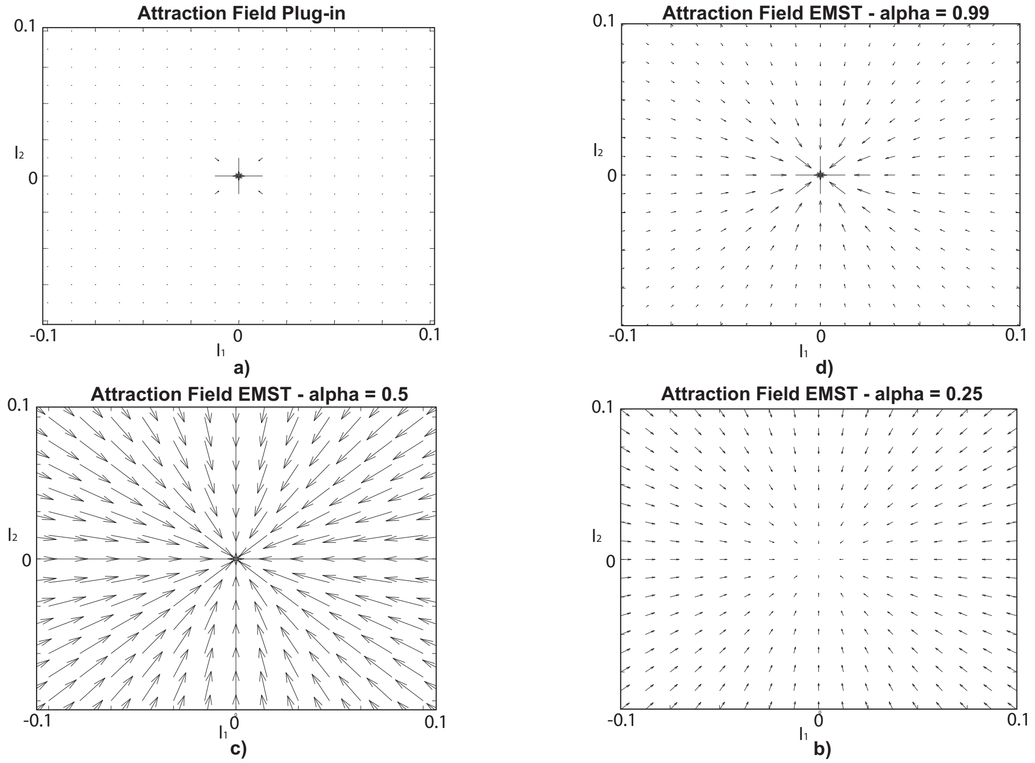


Figure 4.2: Attraction field for a sample at the origin with different entropy estimators and different α values.

arbitrarily scattered values. In an entropic graph estimator, by weighting shorter edges more heavily (with $\alpha > 0.5$), clusters of points drive the algorithm. For a given sample, the entropic graph estimator relies on a small subset of its neighbors, ignoring other samples. This is potentially too aggressive at bad image alignment. On the other hand, in the plug-in estimator, all sample pair attractions are taken into account, and for a given sample the attractions to different clusters are weighted averaged (4.19), where the weights are proportional to the number of samples within the cluster and the inverse of the distance to that cluster. This leads to the following interpretation: the number of samples within a cluster is used as a measure of confidence about these samples being from a correctly aligned region and samples are "pulled into" local high probability regions. Based on this interpretation, we expect the plug-in estimator to be more robust against bad initialization and noise.

At good image alignment, however, relying heavily on the number of samples within a cluster may generate a superfluous attraction to that cluster, merely because it corresponds to a large image structure. Entropic graph estimators avoid this by constraining the attraction generated by a sample to a small number of its neighbors. This allows all clusters, independent of the number of samples, to participate in the fine tuning of the registration result. Thus, we expect the entropic graph estimator to achieve better registration accuracy, especially with high signal-to-noise ratio. This observation is supported by simulations, plotted in Figure 4.3, where the plug-in estimate has a wider basin of attraction and the entropic graph measure has a sharper optimum. Moreover, the lower computational complexity of the entropic graph estimator makes this approach attractive for applications where speed is of concern.

This "sample attractions" interpretation of the registration algorithm provides a justification for the incorporation of the marginal entropy term $H(V)$ (as in mutual information and suggested in (4.3)) with a rich class of transformations, such as the ones that include a zoom component. If the transformation space is rich enough so that the samples can take on any value in the second image, *a **trivial solution** to the registration problem of (4.5) exists. It is when all samples lie on a horizontal line, i.e., when all samples have the same value in the floating image.* In the gradient descent optimization scheme described in Section 4.3, there is no way to explicitly avoid this solution. With rigid-body transformations[5], however, this trivial solution does not exist. This is only a problem with richer transformations, e.g. nonrigid.

[5] constrained by suitable upper and lower limits on the translation and rotation parameters

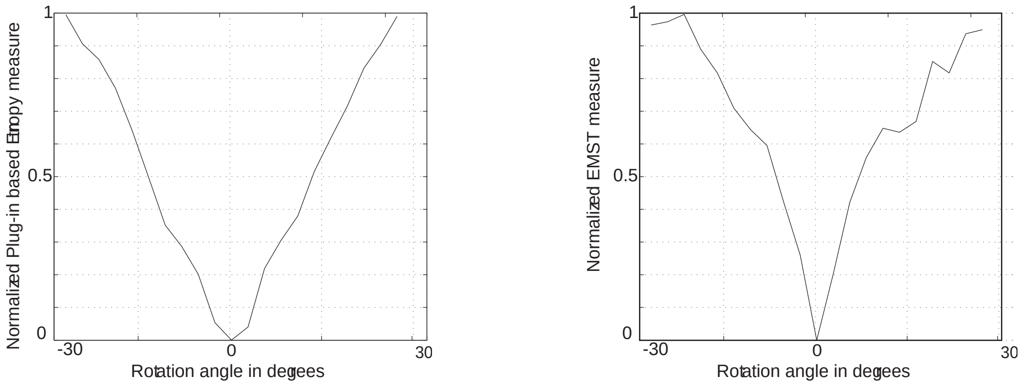


Figure 4.3: Typical profiles of the two entropy estimators with respect to rotation angle. Images shown in Figure 4.5 were used to generate these results.

4.4 Implementation: An EMST-based Rigid Registration Algorithm

In our implementation, we employ spanning trees as the entropic graph family $\mathcal{G}$. The minimal graph is thus the EMST and the misalignment measure is the EMST weight function: $W_\gamma^{MST}(\mathbf{r}) \triangleq W_\gamma^{MST}(\mathcal{S}_\mathbf{r})$. We employ Kruskal's algorithm preceded by a Delaunay triangulation to compute the EMST. The computational complexity of this implementation is $\mathcal{O}(N \log N)$, where N is the number of samples. For details, see Appendix F. Also, note that extension of these ideas to other entropic graphs, e.g. TSP, Steiner tree, nearest neighbor graphs, etc., is also possible.

Figures 4.4 and 4.5 show an image pair and the corresponding EMST of the pixel intensity samples. The profile of the EMST weight for this image pair with respect to rotation angle is also illustrated in Figure 4.3.

In the entropic graph estimator, only with $\alpha \geq 0.5$ is the attractive field's magnitude decreasing as one moves away from the origin (see Figure 4.2). Thus, consistent with our decision to trust clusters, we choose $\alpha \geq 0.5$ in our implementation. However, for $\alpha \geq 0.5$, very close samples undesirably dominate the computation of the function gradient (4.16). Hence, we apply a hard-threshold on $\mathbf{f}_G$ (4.18) and assign a zero value when $\|\mathbf{s}_j - \mathbf{s}_k\|$ is smaller than some small tolerance value.

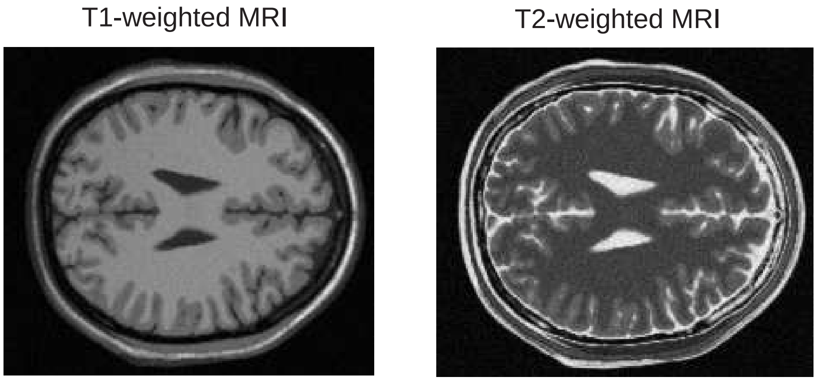


Figure 4.4: Simulated multi-modal images obtained from Brainweb [12].

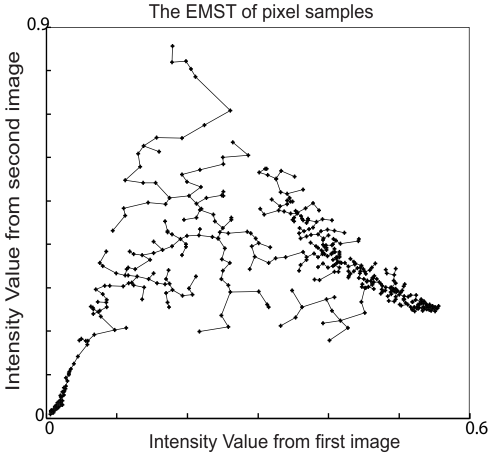


Figure 4.5: The EMST of the intensity sample set.

Simulations suggest that averaging descent directions for different α values yields a smoother profile, making the alignment measure easier to optimize. Recall that, once an EMST is calculated, obtaining descent directions for any α value takes a negligible amount of time, $\mathcal{O}(N)$. Moreover, experimental evidence suggests that α values closer to 1 yield better registration accuracy, whereas smaller α values, i.e., closer to 0.5, yield a wider capture range, making the algorithm more robust against bad initialization. Thus, in our implementation we start the algorithm with $\alpha \approx 0.6$ and gradually increase to ~ 0.9. To minimize the chance of getting trapped in local optima, we employ a multi-resolution pyramid scheme, where the algorithm starts at a coarse resolution and works its way up to the finest resolution. At each level, the initial alignment is obtained from the result of the previous level. In addition, we use quantization within each level to aggregate information. Image intensity values are quantized, initially using a small number of quantization levels. The number of levels is gradually increased. An advantage of this approach is the speed-up of the EMST computation. Moreover, our experiments suggest that the scheme also increases the capture range of the EMST alignment measure.

4.5 Empirical Results

4.5.1 2D Simulations

Figure 4.6 shows a sample pair from the set of 2-D images employed to obtain the registration results summarized in Table 4.2. The second images were artificially created using an intensity mapping, adding i.i.d Gaussian noise and applying a rigid-body geometric transformation consisting of a rotation around the image center and translation along both axes. Thus, ground truth for the alignment was known. The results were obtained by averaging over 100 trials and are the mean square error values with respect to the correct alignment. This experiment compares the multi-modal registration accuracy of four algorithms:

- A1: Plug-in based Renyi entropy estimator with gradient descent optimization
- A2: EMST-based Renyi entropy estimator with descent-based optimization.
- A3: Histogram based normalized mutual information [88] with an implementation of the Nelder-Mead Simplex optimization method [44].

A3 serves as a benchmark, since it is widely accepted as a good entropy-based registration algorithm with acceptable speed and accuracy and reasonable robustness against noise and bad initialization. Note that results for three different cases have been provided:

- Case 1: Good initialization and bad noise: Initial misalignment is a 0-5 pixel translation along each dimension and a 0-5 degrees of rotation. Noise variance is 0.05 times the maximum signal strength.

- Case 2: Moderate initialization and moderate noise: Initial misalignment is a 10-20 pixel translation along each dimension and a 10-15 degrees of rotation. Noise variance is 0.01 times the maximum signal strength.

- Case 3: Bad initialization and small noise: Initial misalignment is a 20-30 pixel translation along each dimension and a 15-20 degrees of rotation. Noise variance is 0.005 times the maximum signal strength.

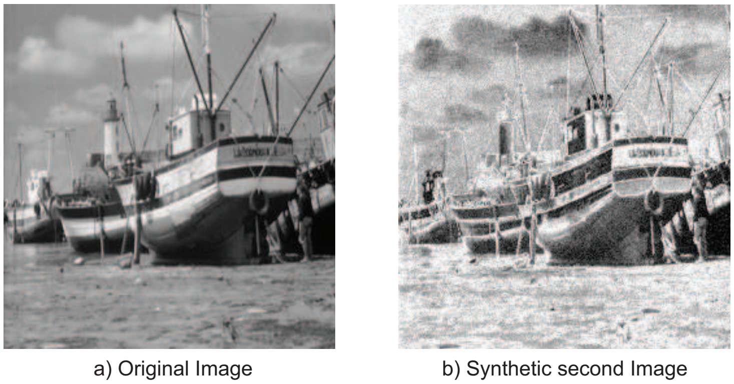
These results confirm our expectation that the plug-in estimator is more robust against bad initialization than the EMST estimator. Note that it is difficult to compare the MSE values of A1 and A2 to the MSE values of A3 since the gradient-descent algorithm employed in the first two algorithms terminated once sub-pixel registration was achieved. The experiments indicate that, once in the basin of attraction, all three algorithms achieve sub-pixel accuracy (see the "Case 1" columns in Table 4.2). Thus, the convergence frequency (CF) values are intended to serve as a measure of the width of the basin of attraction of the corresponding alignment measure. The results suggest that the Rényi based registration algorithms (A1 and A2) have the potential to achieve satisfactory accuracy, and the entropic graph methods yield the fastest run-times.

4.5.2 3D Simulations

In this section, we present results from a 3D rigid-body registration problem. We employ the Brainweb [12] database that consists of simulated MRI volumes ($181 \times 217 \times 181$) of a normal brain at a slice thickness of 1 mm, 3% noise level and 20% RF non-uniformity. Table 4.3 summarizes the registration results using 3D implementa-

	Case 1			Case 2			Case 3		
Algo.	**A1**	**A2**	**A3**	**A1**	**A2**	**A3**	**A1**	**A2**	**A3**
tx	0.68	0.51	0.17	1.12	2.47	0.03	5.34	12.41	0.34
ty	1.85	0.58	0.11	1.54	4.19	0.73	5.78	18.64	0.28
θ	0.70	0.53	0.13	0.87	1.52	0.47	3.82	7.94	0.17
C.F.	100%	100%	100%	72%	55%	99%	41%	32%	100%

Table 4.2: Translation (tx and ty) MSE in pixels, angle (θ) MSE in degrees. Convergence frequency (C.F.) is the percentage of trials where the algorithm achieved sub-pixel accuracy. Average run times (in seconds): 43.1 for **A1**, 6.3 for **A2**, 18.5 for **A3**. Run-times are for Matlab-Mex implementations running on a Pentium IV machine with 512MB RAM.


a) Original Image

b) Synthetic second Image

Figure 4.6: A sample image pair used for the multi-modal registration simulation.

tions of A1 and A2 that employ a stochastic gradient descent optimizer [97], which is a straightforward gradient-descent method performed on a randomly selected subset of the pixels. The stochastic sub-sampling strategy improves run-times and helps avoiding getting trapped in local optima. In both implementations, the number of iterations, number of levels in the multi-resolution pyramid, stopping criterion, step sizes and number of samples were the same. Results show mean square error values (over 100 trials) and suggest that with these data sets, the EMST based algorithm (A2) achieves slightly better accuracy than the plug-in estimator (A1) in a much shorter run-time.

Modality/ Algorithm	x-Trans	y-Trans	z-Trans	Theta	Phi	Omega	Time (sec)
T1 - T2: Ground Truth	20	-10	5	5	-2	7	-
A1: MSE	0.55	0.77	0.70	0.45	0.26	0.36	84.9
A2: MSE	0.36	0.35	0.58	0.22	0.37	0.42	5.96
T1 - PD: Ground Truth	5	-7	3	2	-2	4	
A1: MSE	0.40	0.51	0.40	0.41	0.37	0.39	84.76
A2: MSE	0.46	0.38	0.43	0.24	0.25	0.34	5.81
T2 - PD: Ground Truth	45	5	0	-10	0	5	-
A1: MSE	0.83	0.75	0.55	0.40	0.17	0.30	84.68
A2: MSE	0.26	0.35	0.21	0.17	0.17	0.28	5.96

Table 4.3: 3D registration results using the Brainweb [12] simulated MR volumes. Translation in pixels, angle in degrees. Run-times are for Matlab-Mex implementations running on a Pentium IV machine with 512MB RAM.

4.5.3 3D PET-MR Registration

Here, we present results from a real world application: 3D intra-patient MR-PET rigid registration. Figure 4.7 displays sagittal slices of the two data sets, which were of 128 × 128 × 128 spatial resolution. Figures 4.8 and 4.9 show the volumes before and after registration. Figure 4.10 shows the EMST's computed on pixel intensity values before and after rigid-body registration. The final result was obtained using a MEX/Matlab implementation of the EMST algorithm discussed in Section 4.4. The run-time was approximately 3.5 seconds on a typical Intel machine.

4.5.4 3D MRNeuro Registration

In this section, we present results from a 3D MR Neuro experiment. Figure 4.11 displays sagittal slices of two data sets: a high resolution ($256 \times 256 \times 60$) MR volume and a lower quality MR volume ($128 \times 128 \times 60$) that displays functional information. which were of 128 × 128 × 128 spatial resolution. Figures 4.12 and 4.13 show the volumes before and after EMST-based registration. The run-time was approximately

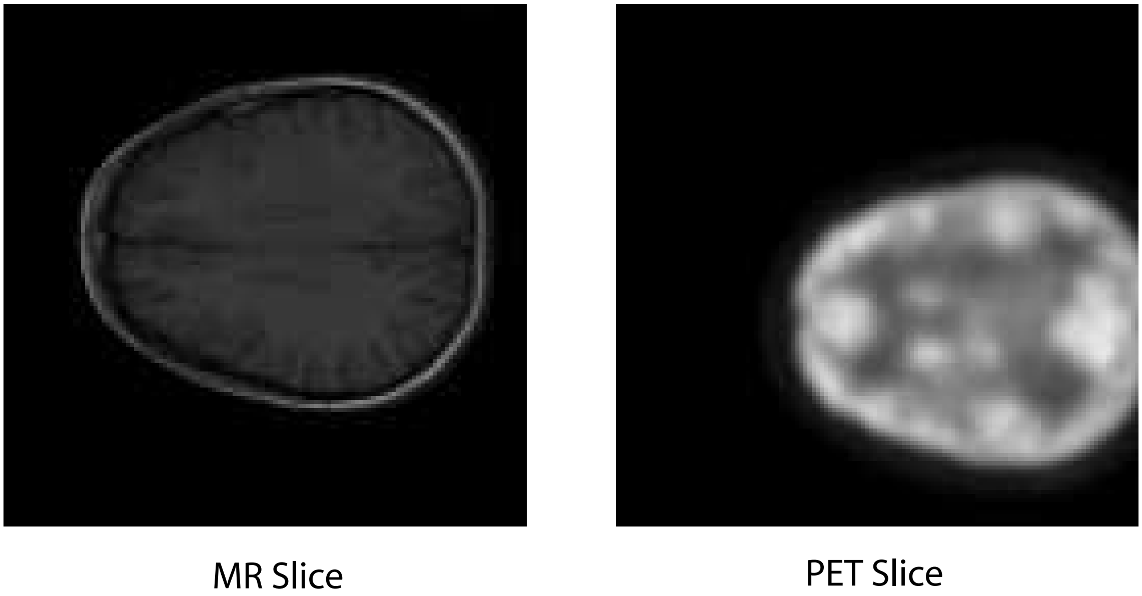


Figure 4.7: Transverse slices of the MR and PET volumes of “*patient 17*”.

2.5 seconds on a typical Intel machine.

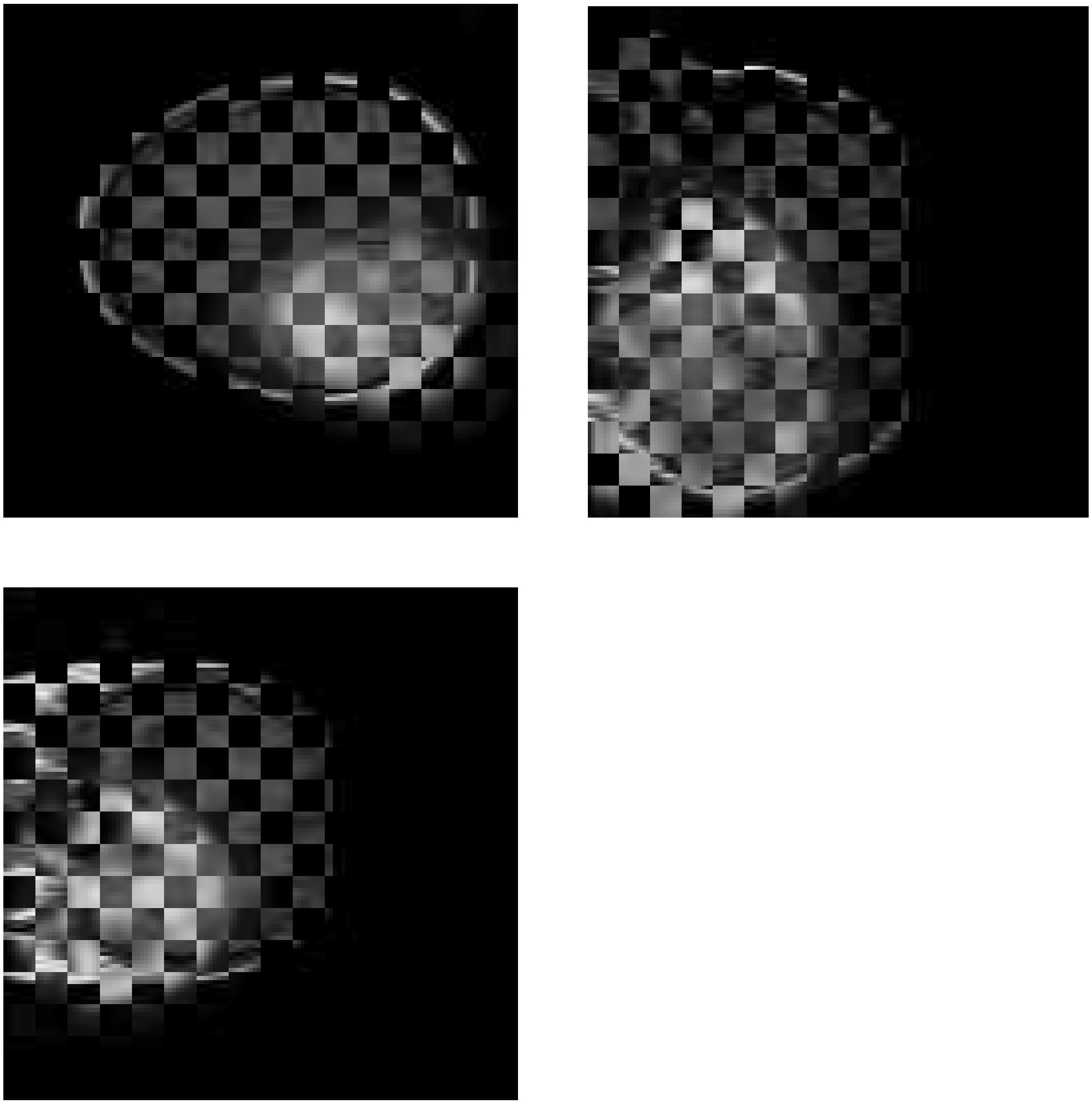


Figure 4.8: Checkerboard representations of the *patient 17* MR and PET data sets at initial alignment (before registration): Transverse, sagittal and coronal views.

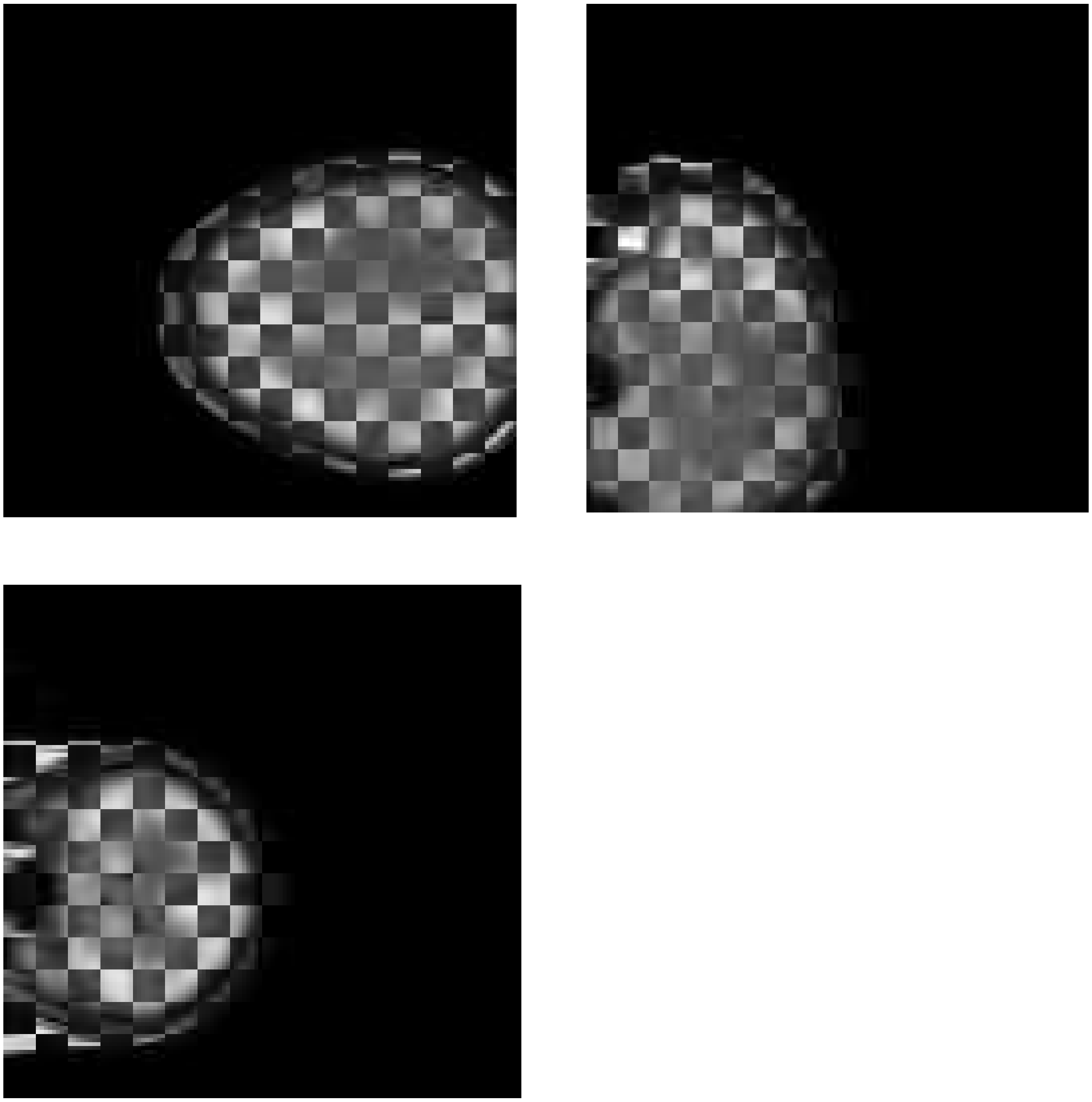


Figure 4.9: Checkerboard representations of the *patient 17* MR and PET data sets after EMST-based rigid-body registration: Transverse, sagittal and coronal views.

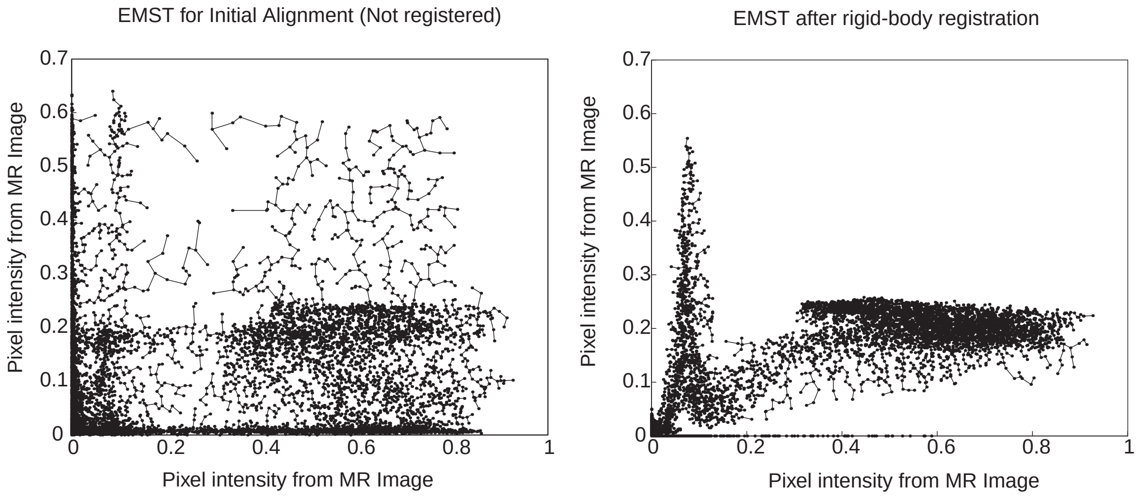
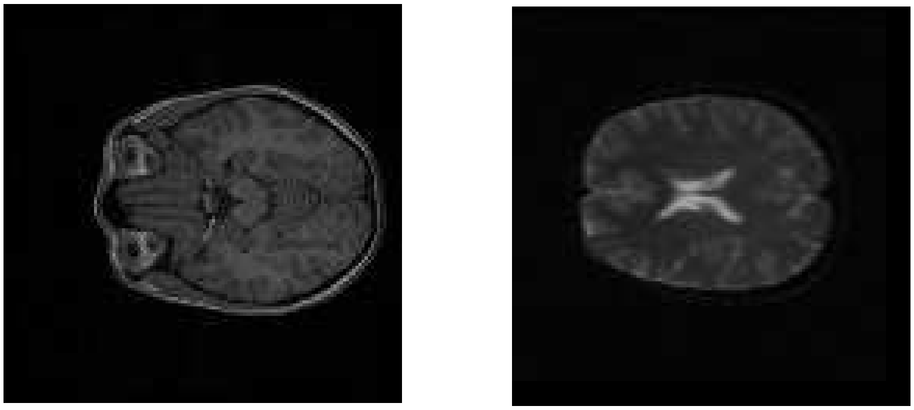


Figure 4.10: Scatter plots and EMST's for before (left) and after (right) EMST-based rigid-body registration of the *patient 17* MR and PET data sets. The average edge length in the EMST's are 0.2145 before registration and 0.1363 after registration.



Hi-res Anatomical MRI

Low-res Lowb MRI

Figure 4.11: Transverse slices of the two different MR "Neuro" volumes.

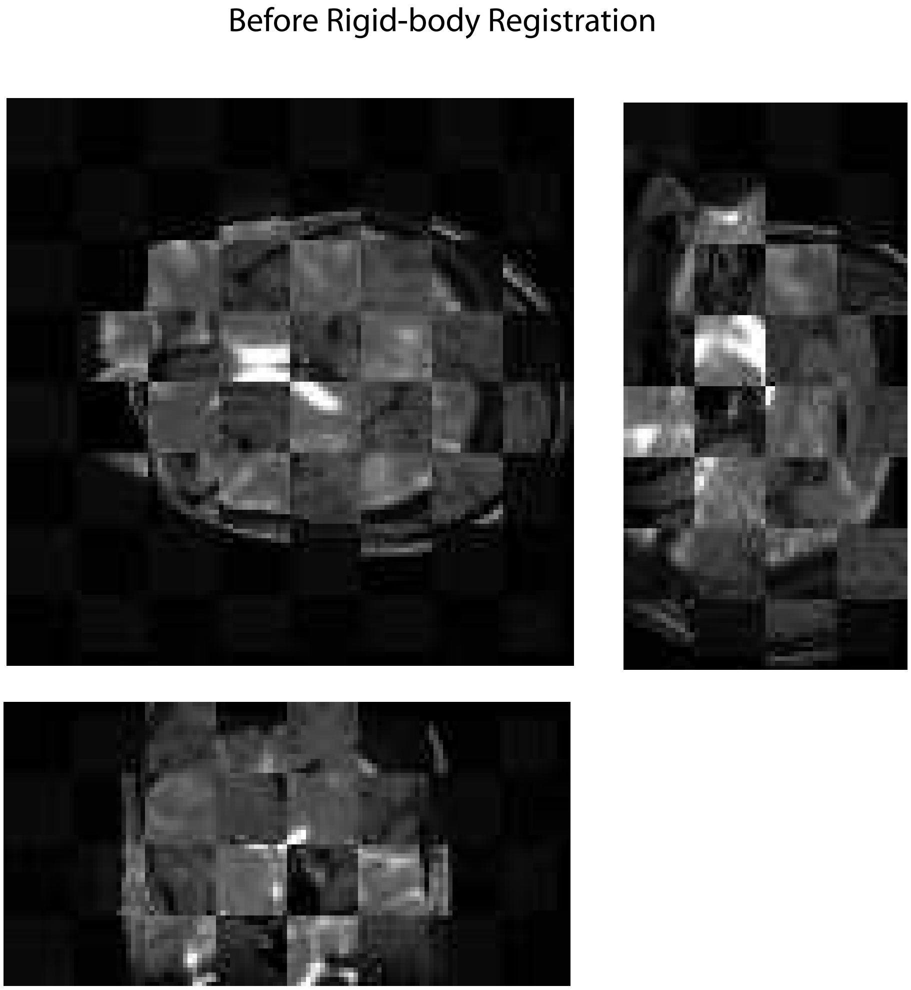


Figure 4.12: Checkerboard representations of the MR “Neuro” data sets at initial alignment (before registration).

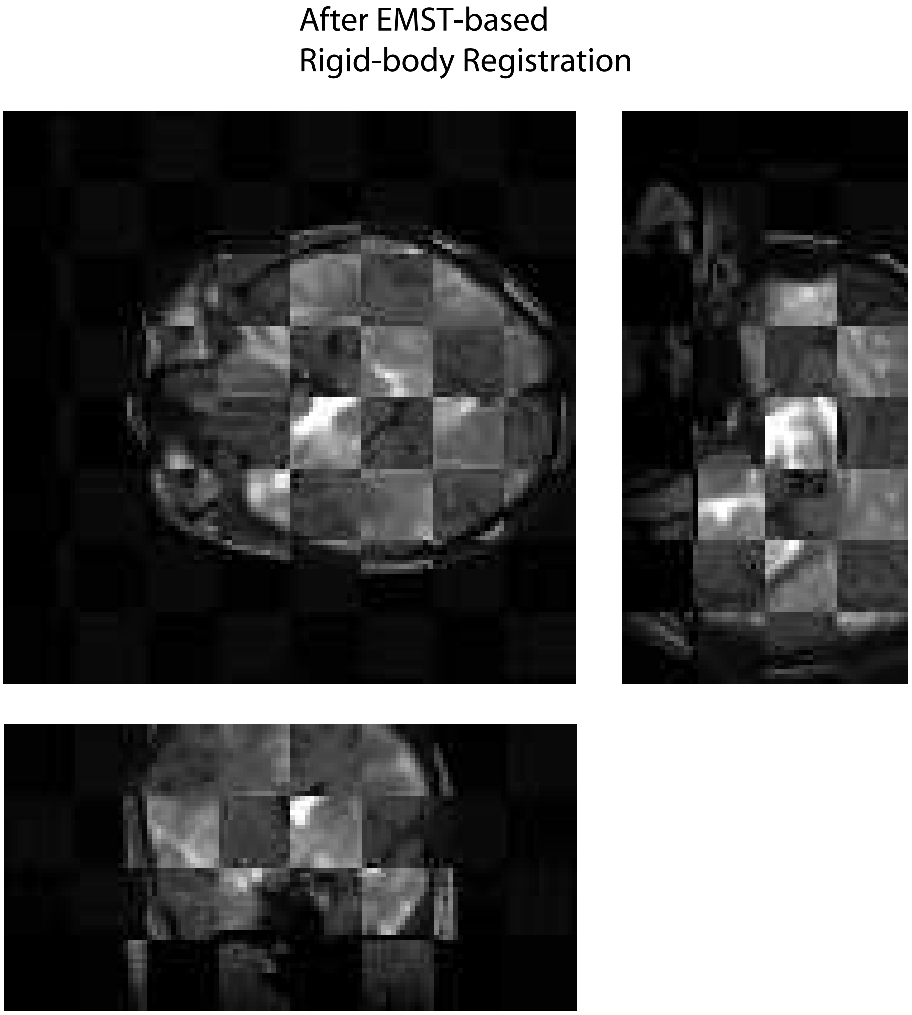


Figure 4.13: Checkerboard representations of the MR "Neuro" data sets after EMST-based rigid-body registration.

Chapter 5

Incorporating Prior Knowledge from Pre-aligned Images

In this chapter, we are interested in the problem of incorporating prior knowledge about the multi-modal relationship from pre-aligned image pairs. In this scenario, we assume that previously aligned images (from the modalities of interest) (also called *training* images) are available to the algorithm that attempts to align a new (*test*) image pair. These pre-aligned images can be manually provided by experts or can be a part of an image sequence the algorithm has already registered. In the following, we examine how we can use this prior knowledge with the entropic graph estimator of entropy. Our main contribution is a method for incorporating prior information in a natural way and with minimal computational overhead into a registration measure based on the Euclidean minimal spanning tree estimate of entropy.

5.1 Introduction and Background

As discussed in previous chapters, images of the same physical structure obtained through different sensing modalities are often assumed to be well modelled through some unknown, yet fixed dependency of the image intensities. For a *registered* image pair, we can usually assume that geometric variations have been corrected for, and that the differences between the two images are mainly due to modality-varying representations of the same structures. The main idea of this chapter is the following: if aligned pairs of images are available, then it should be useful to extract the information about the latter type of (cross-modality) variations between the images and use this information to better the registration algorithm's performance on a new (test) pair of images. In other words, we would like to "learn" the underlying modality relationship.

Consider a situation where we need to register a sequence of multi-modal image pairs (U^k, V^k), $k = 0, 1, 2, \ldots, K$. At time $k-1$ we can assume to have correctly registered the image pairs $(U^j, V^j \circ \Phi_j^*)$ for $j = 0, \ldots, k-1$. Note that, these alignments may convey information about what to expect for the spatial alignment of U^k and V^k, i.e., we may build a prior probability distribution on the geometric transformation space, $p_\Phi(\Phi_k)$, and compute the posterior probability $q(\Phi_k | U^k, V^k, \Phi_1^*, \ldots, \Phi_{k-1}^*)$ to define a similarity measure between U and V. This approach is useful for classification (in a mono-modal setting) and was successfully employed by Miller et al. in [55]. This method typically employs a rich class of transformations with many degrees of freedom. Learning in a high dimensional space requires K to be large.

In the applications we consider in this chapter, however, K is typically small and the problem is multi-modal. In fact, in a lot of cases, we are lucky if we are provided with one or two pairs of correctly aligned images. Moreover the applications typically entail a restricted class of transformations, e.g. rigid-body, affine, etc., and the transformation parameters are considered to be uniformly distributed over a constrained set. In this setting, the modality relationship can be assumed invariant along the sequence and hence information about the modality relationship gained from the prior alignments is potentially useful in the registration of the image pair (U^k, V^k).

An image registration method that does not use prior information gained from previous alignments will be called a *blind* method. If pre-aligned images exist, this case will be referred to as *supervised*. Our goal is to study how prior information obtained from previous registration of multi-modal images can be used to help in the registration problem.

The problem of using prior information to improve multi-modal registration performance was first investigated by Leventon et al. [45]. They propose estimating the prior joint intensity distribution of registered image pairs using training data and then employing a *maximum likelihood* (ML) approach to define the registration measure for new image pairs. Subsequently, Chung et al. [11], proposed an alternative approach in which the quality of registration is determined by the *Kullback-Leibler divergence* between the estimated joint intensity distribution of pre-aligned data and the joint intensity distribution of the new images. Registration is then accomplished by minimizing this K-L divergence. As discussed by Zöllei in [102], the main difference between the ML and K-L divergence approaches is the way they employ the prior distribution to *approximate* the likelihood function. Moreover, it can theoretically be argued that the divergence method [11] yields a monotonic profile of the alignment measure, when close to the global optimum, which is not achieved with the ML method [45]. This suggests that the divergence method is less likely to suffer from getting stuck in local extrema. Both of these studies indicate experimentally that using prior information produces an alignment measure with a wider basin of attraction, making the algorithm more robust to bad initializations, and a registration algorithm that is faster compared to competing methods. Moreover, in [17], Cremers et al. indicate that the employment of prior knowledge improves the robustness and accuracy of the algorithm in nonrigid registration applications.

The main contribution of this chapter is to explore a method to incorporate prior information, as explored in [45] and [11], into the entropic graph based, rigid-body image registration framework, described in the previous chapter.

5.2 Information Divergence Measures

Information divergence measures have been applied to many different domains, including pattern recognition, image and speech processing, machine learning, quantum information theory, graph theory, etc. For a technical treatment of different divergence measures, see [48].

Amongst these, the most popular measure is the Kullback-Leibler divergence[1]. First introduced in [43], K-L divergence, denoted by D, uses Shannon's entropy to define a *non-symmetric* (i.e., directed) distance between *two* probability distributions, p and q:

$$D(p||q) \triangleq \mathbb{E}_p(\log p/q),$$

where $\mathbb{E}_p$ denotes expectation over p. K-L divergence was employed for image indexing and retrieval in [87]. A generalization of the K-L divergence using Rényi's entropy [73] is the so-called α divergence and was employed in [40].

An alternative approach to define divergence measures, namely the Jensen-Shannon (J-S) and Jensen-Rényi (J-R) divergences, relies on the concept of mixing distributions and Jensen's inequality. An advantage of these definitions is that the measures can be defined over multiple distributions and are symmetric.

Definition 5.2.1. *Let $p_1, \ldots, p_k$ be k probability distributions and $\mathbf{w} = (w_1, \ldots, w_k)$ be mixing coefficients, such that $w_j > 0$, $\forall j = 1, \ldots, k$ and $\sum_{j=1}^{k} w_j = 1$. Then, for $\alpha \in (0, 1)$ the J-R divergence is defined as:*

$$J_\alpha^{\mathbf{w}}(p_1, \cdots, p_k) \triangleq H_\alpha(\sum_{j=1}^{k} w_j p_j) - \sum_{j=1}^{k} w_j H_\alpha(p_j),$$

[1] K-L divergence is sometimes called cross-entropy, directed divergence, relative entropy, etc.

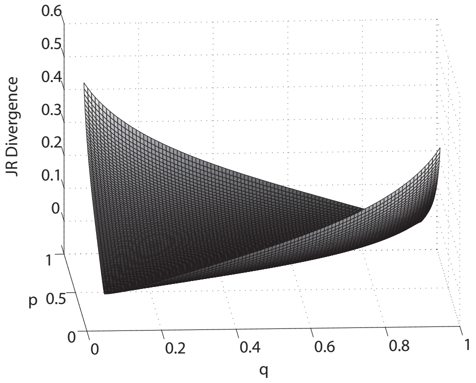


Figure 5.1: J-R Divergence ($\alpha = 0.5$ and $w = 0.5$) between two Bernoulli random variables: p and q.

where H_α is the Rényi entropy.

Using Jensen's inequality, it can be shown that $J_\alpha^{\mathbf{w}}$ is non-negative and achieves zero if and only if all p_j's are equal. Also, the limit of $J_\alpha^{\mathbf{w}}$ as α goes to 1, is equal to the J-S divergence and $J_\alpha^{\mathbf{w}}$ is a convex function of the probability distributions [87]. For example, Figure 5.1 shows the divergence between two Bernoulli distributions with parameters p and q.

5.3 J-R Divergence for Image Registration

Similar to the employment of the K-L divergence in [11], the J-R divergence can be used as a distance between a "correct" distribution and an observed distribution. For example, in [49], $J_\alpha^{\mathbf{w}}$ is simply used to measure the distance between the pixel value distributions of the two images in the overlap region. This alignment measure is suitable for mono-modal applications, where one expects that, at good alignment, the probability of a specific intensity value at a specific location to be similar in the two images. In [100], He et al. provide a thorough analysis of the J-R divergence and show that the divergence measure is maximized for a so-called degenerate set of probability distributions. Inspired by this result, they propose to *maximize* the J-R divergence between the marginal pixel intensity distributions of the floating image within level sets of the reference image. This approach has been demonstrated to yield accurate results in mono-modal applications.

In this thesis, we take a different approach and use the J-R divergence to define a *supervised* misalignment measure for a multi-modal application. This approach is inspired by the employment of the K-L divergence in [11]. For mathematical motivation, we employ the following result, reported in [34], that links the J-R divergence and Bayes decision error:

Let $\mathcal{C} = \{c_1, \ldots, c_k\}$ be a set of k classes, $Y = \{X, C\}$ denote a random variable that takes values on $\mathcal{X} \times \mathcal{C}$ and $f : \mathcal{X} \mapsto \mathcal{C}$ be a classifier. The Bayes' classifier has the minimum misclassification error, $L_B = \min_{f:\mathcal{X}\mapsto\mathcal{C}} P(f(X) \neq C)$. Let $w_i = P(C = c_i)$ be the class probabilities, $\mathbf{w} = (w_1, \ldots, w_k)$, $p_{ij} = P(X = x_j| = c_i)$ be the class-conditional probabilities, and $p_j = (p_{ij}), \forall j = 1 \ldots k$. Based on the original framework presented in [37], Hamza and Krim show that:

$$L_B \leq \frac{H_\alpha(\mathbf{w}) - J_\alpha^{\mathbf{w}}(p_1, p_2, \ldots, p_k)}{2},$$

for $\alpha \in (0, 1]$. In other words, when classifying samples from different distributions, the best performance is achieved when the distributions are maximally distant (as measured by the J-R divergence) to each other.

In a *trained* image registration application, we can assume that there are two "classes" of samples: the ones from the pre-aligned (training) image pair(s) and the ones from the test image pair. Now, consider a scenario where we observed these samples without knowing which image pair they came from. We would expect that the samples would become less distinguishable with better alignment. In other words, the distance (J-R divergence) between the underlying probability distributions should decrease with the quality of alignment. Now, let's make these ideas more formal.

Let $U_t^*(\mathbf{x})$ and $V_t^*(\mathbf{x})$ denote two aligned training images from different modalities;

$U(\mathbf{x})$ and $V(\mathbf{x})$ be two observed images (that are *not* necessarily aligned) from the same respective modalities. Fix a geometric transformation $\Phi : \mathbb{R}^d \mapsto \mathbb{R}^d$. As in previous chapters, assume that each pixel intensity value pair from (U_t^*, V_t^*) and $(U, V \circ \Phi)$ is an independent sample from the distributions p^* and p^Φ, respectively. Then the distance between these distributions is a useful measure to determine the quality of the current alignment. In particular, for $\mathbf{w} = (w, 1-w)$ and $\alpha, w \in (0,1)$:

$$J_\alpha^{\mathbf{w}}(p^\Phi, p^*) = H_\alpha(wp^\Phi + (1-w)p^*) - wH_\alpha(p^\Phi) - (1-w)H_\alpha(p^*) \tag{5.1}$$

can be employed as a *supervised* misalignment measure.

In practice, however, relying heavily on the prior distribution to determine the quality of alignment makes the algorithm's performance sensitive to noise. Also, note the negative marginal entropy term, $H_\alpha(p^\Phi)$, in (5.1). This suggests that in some cases decreasing $J_\alpha^{\mathbf{w}}(p^\Phi, p^*)$ may increase the marginal entropy. Recall that in previous chapters we employed the marginal entropy $H_\alpha(p^\Phi)$ as a *blind* misalignment measure, i.e., to evaluate the quality of alignment based only on the test images. In that framework, the goal of the algorithm was to minimize $H_\alpha(p^\Phi)$.

Based on these observations, we investigate the following hybrid measure that combines the blind and supervised misalignment measures:

$$Q_\alpha(U, V \circ \Phi) = (1-\lambda)J_\alpha^{\mathbf{w}}(p^\Phi, p^*) + \lambda H_\alpha(p^\Phi), \tag{5.2}$$

where $w, \lambda \in (0,1]$ are free parameters. In practice, we choose $w = |\mathcal{S}^\Phi|/(|\mathcal{S}^*|+|\mathcal{S}^\Phi|)$, where $\mathcal{S}^\Phi = \{(U(\mathbf{x}), V \circ \Phi(\mathbf{x}))\}$ and $\mathcal{S}^* = \{(U_t^*(\mathbf{x}), V_t^*(\mathbf{x})\}$; and $\lambda = \frac{w}{1+w}$, where $|.|$ denotes set cardinality. Note that, with these values the weight on the J-R divergence (i.e., the supervised misalignment measure) is proportional to the amount of available prior samples. In other words, if $|\mathcal{S}^*| \gg |\mathcal{S}^\Phi|$, then $\lambda \approx 0$, letting the supervised measure drive the algorithm. On the other hand, if $|\mathcal{S}^*| \ll |\mathcal{S}^\Phi|$, then $\lambda \approx 1$ allowing

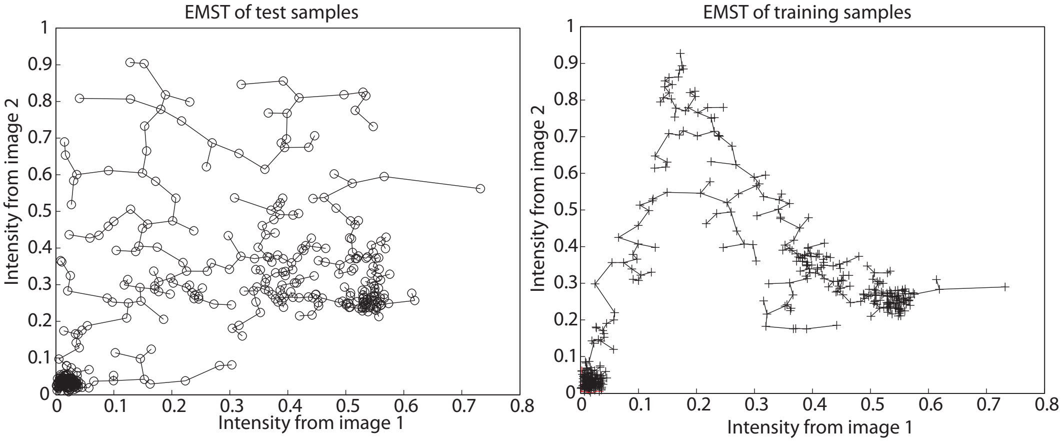


Figure 5.2: EMST's of the test sample set and the training sample set (from pre-aligned images).

the blind measure to drive the algorithm.

Since $H_\alpha(p^*)$ does not depend on the current alignment, it can be removed from the objective function. With the chosen weights, the marginal entropy term cancels and the expression simplifies to:

$$R_\alpha(U, V \circ \Phi) = H_\alpha((1-w)p^* + wp^\Phi). \tag{5.3}$$

5.4 EMST-based Estimate of the Misalignment Measure

Let $\mathcal{S} = \mathcal{S}^* \cup \mathcal{S}^\Phi$. Note that we can assume that the samples in $\mathcal{S}$ are drawn from a mixture distribution equal to $(1-w)p^* + wp^\Phi$, where $w = |\mathcal{S}^\Phi|/(|\mathcal{S}^*| + |\mathcal{S}^\Phi|)$. In the remainder, let's fix $\gamma = 2(1-\alpha)$. Using the entropic spanning graph estimator from Section 3.3.3, $W^*_\gamma(\mathcal{S})$, as defined in (3.7), yields a consistent estimate of (5.3)(e.g., see Figures 5.2 and 5.3, where the pre-aligned images are a simulated t1-t2 MR image pair [12]. For the test (observed) case, the second image was artificially rotated by 10 degrees.). We propose to use $W^*_\gamma(\mathcal{S})$ as a misalignment measure.

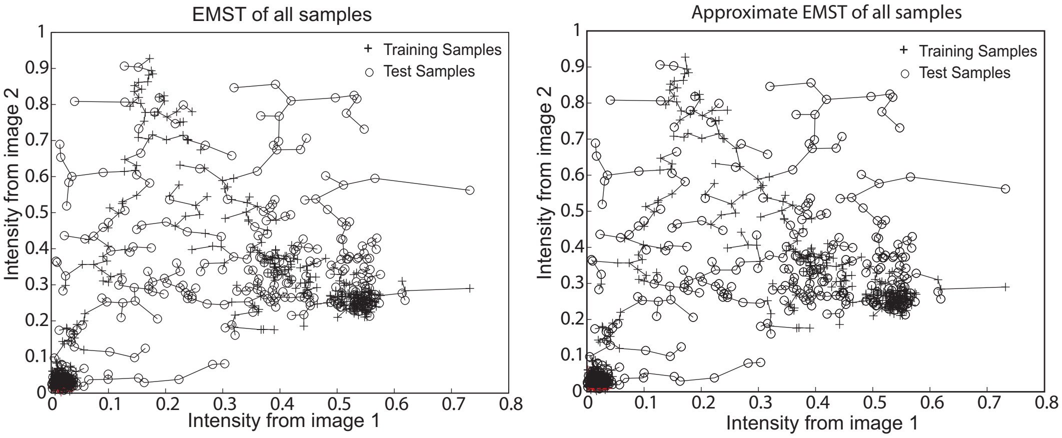


Figure 5.3: EMST and approximate EMST (with $k = 1$, see Section 5.5 for a description) of the union sample set (from observed+pre-aligned images).

Recall that, in Section 4.3.3, we viewed the gradient of the alignment measure as sample pair interactions and in an iterative gradient descent optimization framework, the registration algorithm evolved based on the attractions between samples. In the EMST computed over the union sample set $\mathcal{S}$, the training samples are "stationary," i.e., are independent of the alignment. Thus, they behave like anchors, pulling in the observed samples. This "anchoring effect" makes the algorithm *more robust* to bad initializations. Figure 5.4 shows the effect on the registration function profile (To produce this figure, we used the Brainweb [12] data sets. Training and test images were obtained from different slices of the volume and the test image pair was corrupted by i.i.d Gaussian noise.). It can be seen that the capture range of the misalignment measure has increased when samples from pre-aligned images were used.

Moreover, the proposed supervised misalignment measure, $W^*_\gamma(\mathcal{S})$ is a natural extension of the blind measure investigated in the previous chapter, i.e., $W^*_\gamma(\mathcal{S}^\Phi)$, since $\mathcal{S} = \mathcal{S}^\Phi$ in the absence of training data, i.e., when $\mathcal{S}^* = \emptyset$.

Adding training samples naturally introduces a computational overhead and slows down the algorithm. This overhead can be minimized using an EMST of the training samples, which can be computed off-line. This idea is discussed in the following section.

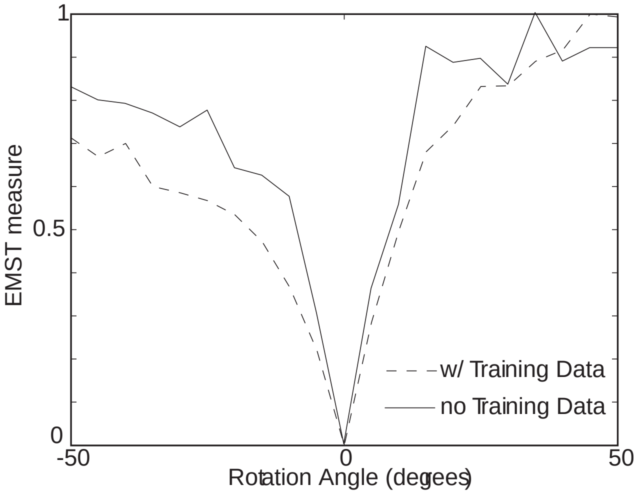


Figure 5.4: EMST measure v.s rotational angle for two cases: with training samples from a pre-aligned image pair and no training samples.

5.5 Computational Issues

The additional computational load of introducing a large set of training samples is important. The following lemma and theorem indicate that an EMST of the training samples, computed off-line, can be used to decrease the computational overhead. Define a minimum spanning forest (MSF) of a graph G as a union of the MST's of the connected components. Note that if G is a complete graph, by Kruskal's algorithm a MSF of G is a MST.

Lemma 5.5.1. *For a graph $G(E, V)$, let $E = E_1 \cup \ldots \cup E_j$ form a partitioning of its edge set. Let F_j denote the edge set of a MSF of $G_j(E_j, V)$. Then there exists a MSF of E (MST if G is complete) such that its edge set $F \subset \cup_i F_i$.*

Proof: Let $e_{12} \in E_j$ denote an edge that connects two vertices v_1 and v_2. If $e_{12} \notin F_j$, then this is the longest edge in a cycle that contains e_{12} and other edges from E_j. Thus, by Kruskal's algorithm this edge cannot be in F.

□

Let T^* and T^Φ be the edges in the EMST's of $\mathcal{S}^*$ and $\mathcal{S}^\Phi$, respectively. Let F^L

be the MSF of the edges that connect samples from $\mathcal{S}^*$ to $\mathcal{S}^\Phi$.

Theorem 5.5.2. *The edges in an EMST of $\mathcal{S}^\Phi \cup \mathcal{S}^*$ are a subset of $T^* \cup T^\Phi \cup F^L$.*

Proof follows from Lemma 5.5.1. □

In our implementation, for large training sets we replace F^L by the set of edges (E_{NN}) that connect each sample in $\mathcal{S}^\Phi$ to its k-nearest neighbors in $\mathcal{S}^*$. This yields a fast approximate EMST algorithm that uses edges in $T^* \cup T^\Phi \cup E_{NN}$. If $|\mathcal{S}^*| >> |\mathcal{S}^\Phi|$, the output tree is usually a good approximation of the complete EMST (see Figure 5.3). Note that, for a fixed observed sample set size $|\mathcal{S}^\Phi|$, a naive computation of the complete EMST is $\mathcal{O}(|\mathcal{S}^*| \log |\mathcal{S}^*|)$ as the training set size grows to infinity. The proposed approximate EMST algorithm that employs a pre-computed EMST reduces this cost to $\mathcal{O}(\log |\mathcal{S}^*|)$.

Lemma 5.5.3. *For a fixed $|\mathcal{S}^\Phi|$, the online computation time of the described approximate EMST is $\mathcal{O}(\log(|\mathcal{S}^*|))$.*

Proof: Computation of T^Φ is $\mathcal{O}(|\mathcal{S}^\Phi| \log |\mathcal{S}^\Phi|)$. Computation of E_{NN} is $\mathcal{O}(|\mathcal{S}^\Phi| \log |\mathcal{S}^*|)$. Given the sorted EMST edge sets T^Φ and T^*, the computation of the final EMST is $\mathcal{O}(|\mathcal{S}^\Phi|)$. The total algorithm is thus $\mathcal{O}(|\mathcal{S}^\Phi| \log |\mathcal{S}^\Phi|) + \mathcal{O}(|\mathcal{S}^\Phi| \log |\mathcal{S}^*|) + \mathcal{O}(|\mathcal{S}^\Phi|)$. □

5.6 Empirical Results

In this section, we provide empirical results for the comparison of the blind and supervised EMST algorithms. Our intention is to illustrate the effect of incorporating prior knowledge into the EMST-based registration framework introduced in this thesis. Thus, we do not benchmark the proposed algorithm against other learning-based registration algorithms, such as [11, 45]. Also, note that our implementation of the EMST-based supervised algorithm gradually omits training samples (ending up with no training samples and only test samples) as the algorithm progresses. This improves the final registration accuracy, while taking advantage of the improved capture range gained through training.

5.6.1 2D Simulations

In our first experiment, we use the simulated natural images and different misalignment scenarios used to produce the results in Section 4.5.1. Since, the second images were synthesized from the original images by applying a (fixed) intensity mapping and corrupting with Gaussian noise, ground truth for alignment was known. Pre-aligned images were obtained with another noise realization. Also, to simulate errors in pre-alignment, we introduced random geometric transformations that did not exceed a one pixel translation (along both axes) and a one degree global rotation.

- Let **A2** designate the registration algorithm that uses an EMST-based Renyi entropy estimator with descent-based optimization and no training samples. This is the same algorithm as A2 in 4.5.1.

- Let **A4** designate the same algorithm as above where the input includes training samples obtained from a pre-aligned image pair.

Results summarized in Table 5.1 provide a confirmation of our expectation that incorporating prior knowledge from pre-aligned images should improve robustness against bad initialization (Recall that Case 3 corresponds to bad initial alignment).

In a different simulation, we wanted to illustrate the effect of incorporating prior knowledge on the final alignment quality under different misalignment conditions. Here, we used the "Bogart" images shown in Figure 5.5. Again, the second image was synthesized from the original first image. This time, the intensity transformation was not one-to-one and was a function of image gradients and intensity values, i.e., violated the common assumption of being a function of pixel intensity values. Figure 5.6

	Case 1		Case 2		Case 3	
Algo.	**A2**	**A4**	**A2**	**A4**	**A2**	**A4**
tx	0.51	0.38	2.47	0.30	12.41	0.33
ty	0.58	0.44	4.19	0.34	18.64	0.38
θ	0.53	0.37	1.52	0.39	7.94	0.44
C.F.	100%	100%	55%	100%	32%	100%

Table 5.1: Translation (tx and ty) MSE in pixels, angle (θ) MSE in degrees. Convergence frequency (C.F.) is the percentage of trials where the algorithm achieved sub-pixel accuracy. Average run times (in seconds): 3.1 for **A2**, 4.2 for **A4**. Runtimes are for Matlab-Mex implementations running on a Pentium IV machine with 512MB RAM.

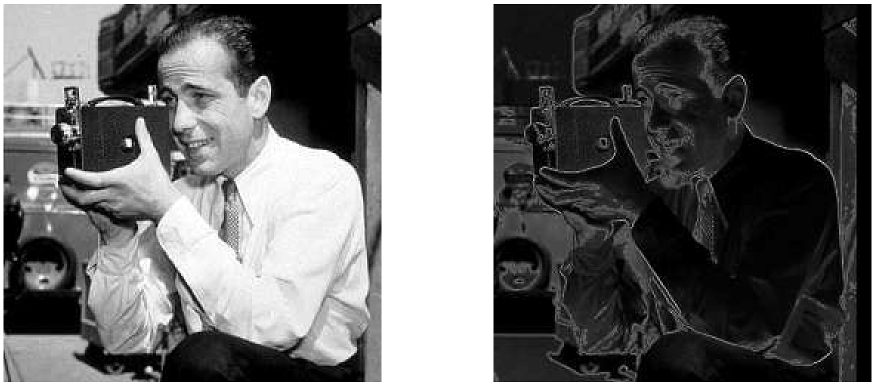


Figure 5.5: The "Bogart" images.

shows the scatter plot for the aligned image pair. Figure 5.7 shows the mean square alignment errors (averaged over 100 trials) versus initial misalignment (translational and rotational). It can be seen that using samples from pre-aligned images yields a better performance with bad initial alignment.

5.6.2 3D Registration

In this section, we present results using the synthetic 3D Brainweb data sets [12], that includes *t*1, *t*2 and *pd* weighted MRI volumes. Figure 5.8 show the sagittal planes of these volumes. To compare the two approaches (blind and supervised), we introduced an initial misalignment by translating or rotating one of the volumes by a relatively large amount. The two algorithms are exactly the same except for

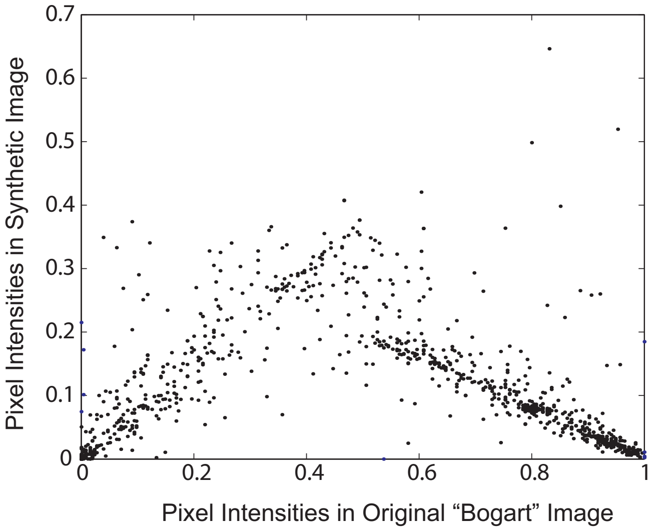


Figure 5.6: Scatter plot of pixel intensity value pairs for correct alignment of the “Bogart” images.

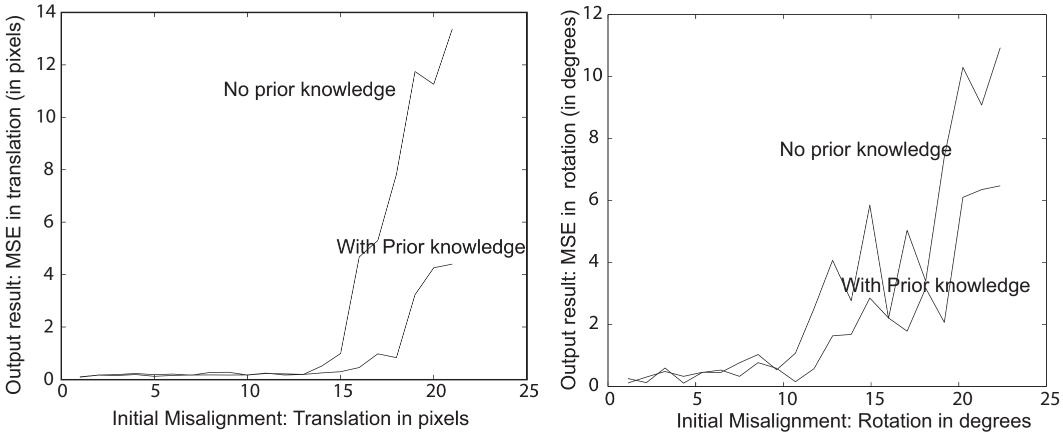


Figure 5.7: Translational and rotational alignment errors (MSE) versus initial misalignment.

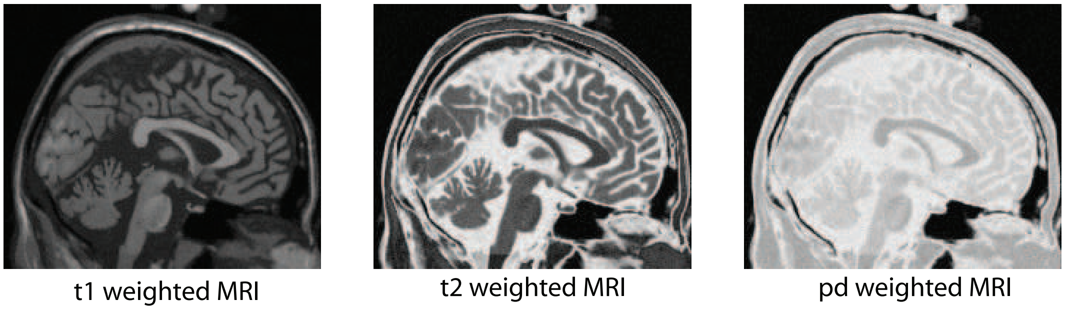


Figure 5.8: Sagittal views of the Brainweb volumes.

their misalignment measures, which are the blind EMST measure (described in the previous chapter) and supervised EMST measure. The pre-aligned images are the original data sets misaligned by some random, yet small rigid body motion (less than one pixel translation and less than one degree rotation in each direction). Table 5.2 displays the "convergence rates" for different cases of initial alignment. The algorithm was declared converged if the final alignment was closer (in absolute difference) than one unit (pixel or degree) to the correct values of *all* transformation parameters. These results were obtained by averaging over 100 trials for each case.

Based on experiments (e.g. see Section 4.5), we know that the EMST-based registration achieves sub-pixel accuracy when transformation parameters are initialized within the capture range of the alignment measure. Thus, the presented convergence rates are intended to serve as a measure of the width of the algorithm's basin of attraction. In all cases, we observe that the supervised algorithm has a higher convergence rate. This supports the claim that *the proposed approach improves the capture range of the EMST based registration algorithm.*

Initial alignment						No training data	With training data
tx	ty	tz	α	β	γ		
pd-t1 registration							
65	0	0	0	0	0	55%	65%
0	65	0	0	0	0	40%	70%
0	0	65	0	0	0	15%	95%
0	0	0	30	0	0	55%	65%
0	0	0	0	25	0	61%	92%
0	0	0	0	0	25	49%	93%
pd-t2 registration							
65	0	0	0	0	0	52%	69%
0	65	0	0	0	0	30%	65%
0	0	65	0	0	0	10%	85%
0	0	0	30	0	0	35%	70%
0	0	0	0	25	0	41%	90%
0	0	0	0	0	25	53%	85%

Table 5.2: Convergence rates of the EMST-based registration algorithm. The employment of training data (i.e., pre-aligned image pairs) as discussed in the text improves robustness against bad initialization.

Chapter 6

Level Set Entropy for Nonrigid Registration

In this chapter, we consider multi-modal applications where global transformation models, e.g. rigid body, are insufficient to capture the geometric variations of interest. In general, we refer to this class of transformations as nonrigid. Specifically, we focus on transformation models that yield a dense deformation field, such as in optical flow techniques. Entropy-based approaches have been investigated for nonrigid registration, but due the computational complexity of the similarity measure and high-dimensional nature of the optimization problem, speed can become a critical issue. In this chapter, we present a linear time nonrigid registration technique that employs a one-dimensional "level set entropy" as a similarity measure within a dense deformation framework regularized by Gaussian smoothing. Similar to the analysis provided in Chapter 2, the proposed measure can be motivated using a maximum likelihood approach. Its main advantage is its flexibility to employ fast and simple entropy estimation techniques. For determining a regularized geometric warp, we show that the Gaussian smoothing technique corresponds to a gradient-descent optimization strategy in a class of smooth and invertible geometric transformations.

Simulations and experimental evidence indicates that level set entropy yields fast and accurate nonrigid registration.

6.1 Nonrigid Registration

In previous chapters, we used entropy-based measures for multi-modal rigid-body registration. We also discussed the straightforward extension of these ideas to other parametric global transformation models, e.g. affine, that include a zooming component. However, in many of today's applications (e.g. multi-subject registration, cardiac motion correction, etc.) the geometric variations across the images can not be adequately described by such global models. Thus, one may desire that the registration algorithm accounts for *local deformations* in the images. We shall use the term nonrigid transformation to refer to any geometric transformation that cannot be captured using a global rigid-body model.

Nonrigid registration is an ill-posed problem: given a sufficiently rich transformation class any image can usually be transformed to be similar to another, a problem analogous to *over-fitting* in machine learning. On the other hand, a conservative transformation class may not achieve the desired alignment accuracy (under-fitting). One approach to circumvent this problem is to employ a restricted class of transformations that provides sufficient flexibility, yet avoids over-fitting, e.g. affine transformations, spline-based parameterized transformations, etc. However, the success of this approach heavily relies on a detailed understanding of the physics of the application. An alternative method is to employ a fairly rich class of transformations, e.g. allowing each pixel to be displaced independently (also known as free-form deformation), but employ an explicit regularization term that reflects our expectations by penalizing “unlikely” transformations.

Speed is an important factor in image registration. Statistical algorithms, e.g.

ones based on entropic alignment measures, can be computationally expensive and the search for the optimum geometric transformation in a rich class of transformations adds an additional computational burden.

The contribution of this chapter is multi-fold: we formulate the registration problem as a joint minimization of a sum of one-dimensional entropies. This formulation provides the flexibility to make stronger, yet realistic, assumptions about the underlying data and yields fast (linear-time) and accurate multi-modal nonrigid registration algorithms. In addition, we examine a fast optical-flow-like method originally derived using an explicit regularizer based on a viscous fluid model [24], and show that this deformation model corresponds to a particular class of smooth transformations.

Section 6.2 motivates the entropy-based misalignment measure. In Sections 6.2, 6.3 and 6.4, we outline our formulation of nonrigid image registration. Details of our implementation and empirical results are provided in Sections 6.5 and 6.6, respectively.

6.1.1 Entropy-based Nonrigid Registration

Motivated by various studies cited in [68], we have employed entropy-based measures to quantify the quality of alignment. The underlying intuition of this approach is that corresponding feature samples (e.g. pixel intensity values, wavelet coefficients, image gradients, etc.) extracted from different images of the same scene become *statistically more dependent* with better alignment. In the sample space, this dependency leads to the clustering of samples, as can be seen in the scatter plots of Figure 2.4.

In registering two images, one image (the reference image) is typically held fixed and the second (floating) image is transformed ("warped") so that it is aligned with the reference. In this setting[1], for each feature sample only the component that corresponds to the floating image value varies as the floating image is warped. When the

[1] Assuming we use a fixed set of sampling locations

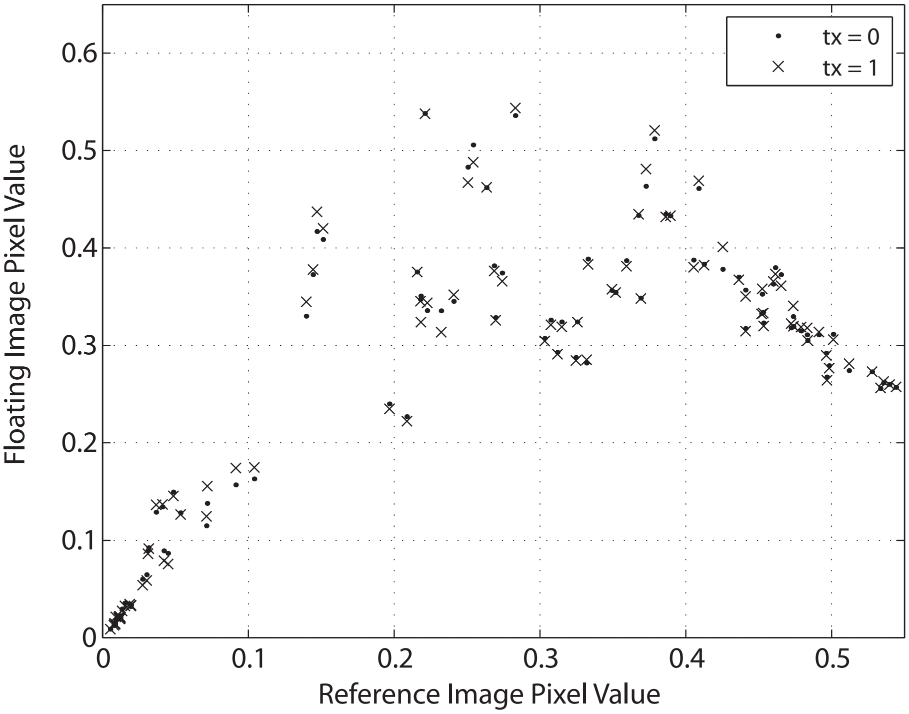


Figure 6.1: Scatter plot of pixel intensity samples from a pair (reference and floating) of images before ($t_x = 0$) and after ($t_x = 1$) the floating image is translated by one pixel. Notice that for the two cases the sample values vary in their second coordinate values only.

extracted features are pixel intensities, this manifests itself as the vertical movement of the samples in the intensity scatter plot. See for example Figure 6.1, where a small number of samples are shown before and after a 1-pixel translation of the floating image.

To our knowledge, this constraint on the sample value change has not been explicitly exploited in prior entropy-based registration algorithms. One method to exploit it is to estimate the one-dimensional entropy of samples that have a fixed intensity value in the reference image. This approach is similar to the so-called “congealing” technique [53, 104] on level sets of the reference image. This gives us the flexibility to make strong assumptions about the samples within each level set and in turn yields faster and possibly more accurate algorithms. For instance, if we know that the cross-modality intensity relationship is one-to-one at alignment, i.e., we expect samples from a level set of the reference image to cluster around one value at good alignment, then a unimodal distribution, e.g. a Gaussian, can model the underlying distribution in each level set. Moreover, by treating each level set independently, we are relaxing the assumption, implicit in most algorithms, that the underlying cross-modality relationship is continuous.

Nonrigid image transformations can be parametric or nonparametric. Parametric transformations employ a parameterized transformation space, e.g. affine transformation, thin-plate splines, B-splines, etc. The goal is to determine the set of parameter values that optimize a fixed alignment measure. This technique has been used with entropy-based measures in [80, 76, 47, 64]. The advantage of parameterized techniques is that the dimensionality of the problem is relatively low and hence a robust optimization is possible. However, in some applications it is not clear how to select a natural parameterized transformation space. Moreover, these approaches often require quite a few major design decisions, e.g. number of control points in the B-spline, that have an important influence on the final result and thus have to be fine-tuned for each application.

In a nonparametric approach (also referred to as optical-flow-like deformation, dense deformation, etc.), each image pixel is transformed independently. To circumvent the ill-posedness of the problem and to incorporate prior knowledge about the transformation, one can employ a gradient-descent-like time marching scheme to minimize a global functional of the geometric transformation. This functional consists of two terms: the alignment measure and an external regularization term that reflects our expectations by penalizing unlikely transformations. Depending on the application, various energy functionals have been proposed in the literature. While most of these are inspired by physical models, e.g. elasticity, viscous fluid and diffusion models [56], others employ a Bayesian approach with a prior distribution model, e.g. Brownian warps [61]. An alternative strategy, also motivated by physical models, is an iterative scheme where a "rough" warp field obtained from the gradient of the alignment measure is projected onto a known function space. This projection is done by spatial smoothing [65, 18, 38]. This approach has yielded fast nonrigid registration algorithms [24].

In the following, we investigate a fast smoothing-based optical-flow-like method with the proposed "one-dimensional" entropy-based alignment measure. As discussed in Section 6.4, this can be motivated using a gradient descent optimization strategy in a family of smooth transformations.

6.2 Level Set Entropy as a Misalignment Measure

Let $U(\mathbf{x})$ and $V(\mathbf{x})$ be images, where $\mathbf{x} \in \mathbb{R}^d$ and $d = 2$ or 3. Let $\Omega \subset \mathbb{Z}^d$ be a finite region of interest. Write the warp function in the form $\Phi(\mathbf{x}) = (\mathbf{Id}+h)(\mathbf{x}) = \mathbf{x}+h(\mathbf{x})$, where $\mathbf{Id}$ is the identity transformation and $h : \mathbb{R}^d \mapsto \mathbb{R}^d$. We assume that the warp is applied to the second image to produce $V(\Phi(\mathbf{x}))$.

Assuming that pixel intensity values $U(\mathbf{x})$ and $V(\Phi(\mathbf{x}))$ are independent samples from p_U and p_V and using a maximum likelihood approach, as in Section 2.3.1, we derived the conditional entropy $H(V(\Phi(\mathbf{x}))|U(\mathbf{x}))$ as a misalignment measure. Now, let's define the level set entropy of the image $V \circ \Phi$ for $U = u$ as:

$$H(V \circ \Phi(\mathbf{x})|U(\mathbf{x}) = u) \triangleq -\int \log(p_V(V \circ \Phi(\mathbf{x})|U(\mathbf{x}) = u))p_V(V \circ \Phi(\mathbf{x})|U(\mathbf{x}) = u)dV. \tag{6.1}$$

Note that $H(V \circ \Phi(\mathbf{x})|U(\mathbf{x})) = \mathbb{E}(H(V \circ \Phi(\mathbf{x})|U(\mathbf{x}) = U))$, where the expectation $\mathbb{E}$ is over the distribution p_U. Given an estimator $\hat{H}(V \circ \Phi(\mathbf{x})|U(\mathbf{x}) = u)$ for (6.1), a sample mean estimate of $H(V \circ \Phi(\mathbf{x})|U(\mathbf{x}))$ is:

$$\hat{H}(V \circ \Phi(\mathbf{x})|U(\mathbf{x})) = \frac{1}{N}\sum_u N_u \hat{H}(V \circ \Phi(\mathbf{x})|U(\mathbf{x}) = u), \tag{6.2}$$

where $N = |\Omega|$ and $N_u = |\Omega_u| = |\{\mathbf{x} : \mathbf{x} \in \Omega, U(\mathbf{x}) = u\}|$. Also, by the law of large numbers $p_U(u) \approx N_u/N$. In the following, we explore (6.2) as a misalignment measure. This entails the estimation of the one dimensional entropic measure (6.1).

At this point, it is important to emphasize that the *trivial solution* discussed in Section 4.3.3, is not of immediate concern, since in the proposed level set entropy framework, inter-sample attractions are only effective within each level set. In other words, samples that have the same fixed image value attract each other. Thus, even if all samples were allowed to move independently, it is highly unlikely that they would line up horizontally and take on the same value in the second image.

6.3 Entropy Estimation

To estimate the one-dimensional level set entropy, we will employ "plug-in" entropy estimators (see Chapter 3 for a more detailed overview). In this approach, a density estimate is inserted into the entropy formula yielding an entropy estimate. *Parametric* density estimators employ a family of parameterized densities (e.g. a Gaussian density, mixture densities, etc.) and estimate the parameter values based on observations. *Nonparametric* methods, on the other hand, let the data (e.g. histogram) determine the "shape" of estimated density.

Let $\mathcal{S} = \{s_1, \ldots, s_M\}$ be M independent one-dimensional samples of a continuous density p_S. If we assume p_S is a Gaussian density, a parametric plug-in estimate of the corresponding entropy is:

$$\hat{H}_G(S) = \frac{1}{2}\ln(2\pi \mathrm{e}\hat{\sigma}^2), \tag{6.3}$$

where $\hat{\sigma}^2 = \sum_{s_i \in \mathcal{S}} (s_i - \bar{s})^2/M$ and $\bar{s} = \sum_{s_i \in \mathcal{S}} s_i/M$ are the maximum likelihood estimates of the variance and mean, respectively. Note that, parametric estimators for other types of distributions can also be employed. See, for example, Section 3.4.

A nonparametric estimate of the density, given by the Parzen window method, is:

$$\hat{p}_S(x) = \frac{1}{M} \sum_{s \in \mathcal{S}} g_\sigma(x - s), \tag{6.4}$$

where $g_\sigma(\cdot)$ is the one-dimensional zero-mean Gaussian with a variance σ^2. Plugging (6.4) into the entropy formula yields a nonparametric plug-in estimate of the entropy:

$$\hat{H}_P(S) = -\int_{\mathbb{R}} \hat{p}_S(x) \log \hat{p}_S(x) dx. \tag{6.5}$$

6.4 Warp Field

As in [61], we assume that the geometric mapping $\Phi(\mathbf{x})$ is a concatenation of "smooth" invertible mappings. Let $\Phi_j^\sigma(\mathbf{x})$ be an invertible function that can be expressed as:

$$\Phi_j^\sigma(\mathbf{x}_0) \triangleq \mathbf{x}_0 + \sum_{\mathbf{x} \in \Omega} \mathbf{r}_j(\mathbf{x}) \bar{G}_\sigma(\mathbf{x}_0 - \mathbf{x}),$$

for some *vector field* $\mathbf{r}_j : \Omega \mapsto \mathbb{R}^d$ and normalized discrete Gaussian filter

$$\bar{G}_\sigma(\cdot) = G_\sigma(\cdot) / \sum_{\mathbf{x} \in \mathbb{Z}^d} G_\sigma(\mathbf{x}),$$

where $G_\sigma(\cdot)$ is the d-dimensional zero-mean Gaussian with covariance matrix $\Sigma = \sigma^2 I$. Gaussian smoothing can be theoretically motivated by a viscous fluid model, as shown in [24]. Now, for a $L \in \mathbb{Z}^+$ we assume that the geometric mapping Φ belongs to a family $\mathcal{W}(L, \sigma)$ consisting of functions in the following form:

$$\Phi(\mathbf{x}) = \Phi_1^\sigma(\mathbf{x}) \circ \ldots \circ \Phi_L^\sigma(\mathbf{x}). \tag{6.6}$$

By the Inverse Function Theorem and assuming the mapping is orientation-preserving, i.e., the image is not flipped, $\Phi(\cdot)$ is invertible if it has a positive Jacobian determinant, i.e., $\det(J_\Phi) > 0$ at every point $\mathbf{x} \in \mathbb{R}^d$, where $J_\Phi(i,j) = \partial\Phi^j/\partial x^i$. By the chain rule and (6.6):

$$J_\Phi = \prod_{j=1}^{L} J_{\Phi_j}.$$

It is easy to see that if each Φ_j has a positive Jacobian, then each Φ_j is invertible and thus Φ is invertible with:

$$\Phi^{-1} = \Phi_L^{-1} \circ \ldots \circ \Phi_1^{-1}. \tag{6.7}$$

Note:

$$J_{\Phi_j}(\mathbf{x}_0) = I + \sum_{\mathbf{x}\in\Omega} \nabla\bar{G}_\sigma(\mathbf{x}_0 - \mathbf{x})^T \mathbf{r}_j(\mathbf{x}), \tag{6.8}$$

where I is the identity matrix, $\nabla\bar{G}_\sigma$ is the gradient of $\bar{G}$ and M^T denotes the transpose of M.

Using the level set entropy misalignment measure, we formulate image registration as:

$$\Phi^* = \underset{\Phi\in\mathcal{W}(L,\sigma)}{\operatorname{argmin}}\ \hat{H}(V \circ \Phi | U), \tag{6.9}$$

where $\hat{H}(V \circ \Phi|U)$ is defined in (6.2). A suboptimal solution to (6.9) is $\Phi^{**} = \Phi_1^* \circ \ldots \circ \Phi_L^*$, where

$$\Phi_j^* = \underset{\Phi_j\in\mathcal{W}(1,\sigma)}{\operatorname{argmin}}\ \hat{H}(V \circ \Phi_1^* \circ \ldots \circ \Phi_{j-1}^* \circ \Phi_j) \tag{6.10}$$

and $\Phi_0^* = \mathbf{Id}$. This is similar to the *re-gridding* approach used in [10]. Let $\mathcal{R}(\sigma) = \{\mathbf{r} : \Omega \mapsto \mathbb{R}^d \text{ s.t } \Phi_\mathbf{r}^\sigma \text{ is invertible in } \Omega\}$. Define $V^j = V \circ \Phi_1^* \circ \ldots \circ \Phi_j^*$ for $j = 1, \ldots, L$ and:

$$\mathbf{r}_j^* \triangleq \underset{\mathbf{r}_j\in\mathcal{R}^\sigma}{\operatorname{argmin}} \hat{H}(V^{j-1} \circ \Phi_j | U). \tag{6.11}$$

Then, by definition:

$$\Phi_j^*(\mathbf{x}_0) = \mathbf{x}_0 + \sum_{\mathbf{x}\in\Omega} \mathbf{r}_j^*(\mathbf{x})\bar{G}_\sigma(\mathbf{x}_0 - \mathbf{x}).$$

As in [38], employing the first variation of $\hat{H}$ within a "gradient descent-like" optimization strategy to solve (6.11) yields the following iterative algorithm:

$$\begin{aligned}
\mathbf{r}_j^0 &= 0, \Phi_j^0 = \mathbf{Id} \\
\mathbf{r}_j^{t+1}(\mathbf{x}_0) &= \mathbf{r}_j^t(\mathbf{x}_0) - \lambda \sum_{\mathbf{x}\in\Omega} F(\mathbf{x})\bar{G}_\sigma(\mathbf{x}_0 - \mathbf{x}), \qquad (6.12) \\
\Phi_j^{t+1}(\mathbf{x}_0) &= \mathbf{x}_0 + \sum_{\mathbf{x}\in\Omega} \mathbf{r}_j^{t+1}(\mathbf{x})\bar{G}_\sigma(\mathbf{x}_0 - \mathbf{x}), \qquad (6.13)
\end{aligned}$$

where

$$F(\mathbf{x}) \triangleq \frac{\partial \hat{H}(V^{j-1} \circ \Phi_{\mathbf{r}_j^t} | U)}{\partial V^{j-1}(\Phi_{\mathbf{r}_j^t}(\mathbf{x}))} \nabla V^{j-1}(\Phi_{\mathbf{r}_j^t}(\mathbf{x})) \qquad (6.14)$$

is the gradient field of the entropy estimate, $\lambda > 0$ is a step size and ∇I is the gradient image of I. Combining Equations (6.12) and (6.13), an equivalent algorithm is:

$$\begin{aligned}
\Phi_j^0 &= \mathbf{Id} \\
\Phi_j^{t+1}(\mathbf{x}_0) &= \Phi_j^t(\mathbf{x}_0) - \lambda \sum_{\mathbf{x}\in\Omega} F(\mathbf{x}) \qquad (6.15) \\
&\quad * \bar{G}_{\sqrt{2}\sigma}(\mathbf{x}_0 - \mathbf{x}) W_{\sigma/\sqrt{2}}((\mathbf{x}_0 + \mathbf{x})/2),
\end{aligned}$$

where $W_\sigma(\mathbf{x}) \triangleq \sum_{\mathbf{y}\in\Omega} \bar{G}_\sigma(\mathbf{y} - \mathbf{x})$. Note $W_\sigma(\mathbf{x}) \leq 1$, and $W_\sigma(\mathbf{x}) = 1, \forall \mathbf{x} \in \mathbb{Z}^d$, when $\Omega = \mathbb{Z}^d$. In practice, we use this upper bound to speed up the algorithm. For small σ, this is a good approximation away from the boundary of Ω.

6.5 Algorithm

6.5.1 Gradient Field of Entropy Estimate

To compute the gradient field of the entropy estimates (6.3) and (6.5), we need to compute the derivative of the entropy estimates with respect to a sample value. For the Gaussian parametric estimate, the derivative is:

$$\frac{\partial \hat{H}_G(S)}{\partial s_j} = \frac{1}{M}\frac{s_j - \bar{s}}{\hat{\sigma}^2}, \tag{6.16}$$

and for the Parzen-window based estimate the derivative is:

$$\frac{\partial \hat{H}_P(S)}{\partial s_j} = -\frac{1}{M}\int_{\mathbb{R}} g'_\sigma(s_j - x)\log \hat{p}_S(x)dx, \tag{6.17}$$

where g' is the derivative of the Gaussian. Using these, we can easily compute the gradient field expressions (6.14) for the corresponding entropy estimates.

6.5.2 Implementation

The proposed algorithm combines all three components described in the previous sections. The conditional entropy of the pixel intensity values is estimated using (6.2). We employ two different entropy estimators, parametric (6.3) (*Algorithm 1*) and nonparametric (6.5) (*Algorithm 2*), to estimate the one-dimensional level set entropy (6.1). The level sets are computed once at the beginning of the algorithm by determining regions (pixel locations) with the same quantized reference image value. We used 10-50 quantization levels. Note that the computation of the entropy gradients are linear time, i.e., $\mathcal{O}(N)$, where N is the total number of pixels. We employ a blurred histogram (typically with 20-30 bins) to quickly compute a density estimate (6.4). Also, the integral of Equation (6.17) is approximated using a finite sum over the histogram bins. One trick we employed to speed-up the algorithm (especially with high-resolution 3D data sets) is to use a subset of the pixels to compute the sample means $\bar{s}$ and variances $\hat{\sigma}^2$ in Algorithm 1 or the density estimate $\hat{p}_S$ (6.4) in Algorithm 2.

The optimization component is an iterative scheme, where at each iteration the floating image is warped by smoothing the gradient field of the entropy estimate with a finite length normalized Gaussian filter, as in Equation (6.15). In practice, for 2D images of size 128 × 128 we used a 10 × 10 Gaussian filter, and a step size satisfying $\lambda\|F\| < 0.3$, where $\|F\| \triangleq \max_{\mathbf{x}\in\Omega} |F(\mathbf{x})|$ is the maximum gradient field magnitude, and $\sigma = 5$. To avoid singularities in the warp, one can perform a re-gridding whenever the Jacobian (6.8) comes too close to zero, i.e. $\min_{\mathbf{x}\in\Omega} J_\Phi(\mathbf{x}) < 0.1$. In practice, we found that re-gridding after each iteration with the given step size sped up the algorithm dramatically, while producing good results. As widely done in image registration, we constructed a standard Gaussian multi-resolution pyramid (with 3-4 levels) to improve the accuracy and speed of the algorithm. At non-integer locations we used bilinear interpolation to compute intensity values.

6.6 Empirical Results

Validating a nonrigid image registration algorithm is a difficult task. An important indicator of the quality is its performance with simulated examples. Yet, we believe that the real value of an algorithm can only be revealed within an application. The following results will thus serve only as a preliminary evaluation of our proposed algorithm and a confirmation of our expectations. Quantitative results are presented using simulated data, i.e., with ground truth available. Qualitative evidence is also given based on visual inspection of a real-world application.

6.6.1 Simulations

We employ the Brainweb images [12]. In each simulation, the floating image is generated by applying a known warp field to one of the images. The goal of the registration algorithm is to recover this warp, which was created using a thin-plate spline model [5] (different than the proposed algorithms' deformation model), and is invertible and smooth.

Here, we present two cases: 1) t1-t2, and 2) t2-pd registration. In the first case, the two modalities have a one-to-one relationship at perfect alignment, as can be seen in Figure 6.2-a. Thus, we expect the parameterized entropy estimator (based on a Gaussian model), i.e., Algorithm 1, to perform well. In the second case (Figure 6.2-b), the cross-modality relationship is not one-to-one, hence the Gaussian model is too restrictive and accuracy suffers. Difference images are provided in Figures 6.4 and 6.5 for pre-registration and post-registration. Figure 6.3 displays the original and recovered warp fields for the *t1-t2* experiment using deformed grids. Table 6.1 summarizes the quantitative results. Run times are for a Matlab implementation running on a Pentium 4 machine with a 512MB RAM.

	intensity MSE	grid MSE	time (sec.)
t1-t2 Registration			
Artificial warp	0.0156	15.11	-
Algorithm 1	0.0029	2.81	9.09
Algorithm 2	0.0037	3.09	10.17
pd-t2 Registration			
Artificial warp	0.0181	16.32	-
Algorithm 1	0.0079	4.98	9.01
Algorithm 2	0.0041	3.33	10.2

Table 6.1: Simulation results for *t1-t2* and *pd-t2* MR registration. Intensity MSE: Mean square difference between the intensity values (in [0,1]) of original and warped floating image. Grid MSE: Mean square difference between deformed grid and ground truth (in pixel coordinates).

In Table 6.1, the first column (intensity MSE) attempts to evaluate the result based on pixel intensity values, i.e., how similar the recovered floating image is compared with its original. The values suggest that both algorithms do a good job in making the images look similar, but this measure ignores the actual warp. The second column, on other hand, measures the discrepancy between the grid warped consecutively using the synthetic and recovered warps, and a uniform grid. Under perfect conditions, the consecutively warped grid should be uniform. Thus, these values indicate the algorithm's success in recovering a warp. Note that, after the synthetic warp is applied, the average distortion between the original (uniform) grid and deformed grid is about 15-16 pixels. Both algorithms bring down these values to 2-4 pixels. As expected, Algorithm 1 performs better in the first case. Algorithm 2 achieves better results in the second case, where there is no one-to-one relationship between pixel intensity values in the two images at perfect alignment.

6.6.2 3D Experiment

In this section, we present results from a mono-modal multi-subject application. Figures 6.7 and 6.6 show the MR volumes (of resolution $128 \times 128 \times 128$) of two subjects. Here, ground truth and an objective measure of alignment quality is not available. One approach to determine registration accuracy is through visual inspection.

To achieve the final alignment results, we employed a three step strategy. As commonly done in brain imaging, in the first step we pre-processed the images to extract "non-brain" regions (e.g. the skull). For this, we utilized a "Brain Extraction Toolkit" developed by Smith [86] within the MRIcro environment [77]. Figure 6.8 displays the "stripped" brains at initial alignment. In step 2, we brought the two brains into rigid alignment using the EMST algorithm described in Chapter 4. A MEX/Matlab implementation running on 512MB Intel machine took about 7 seconds to complete this step. Figure 6.9 shows the volumes after rigid registration. Finally, in step 3, we employed the level set entropy-based nonrigid registration algorithm investigated in this chapter and described as Algorithm 1 in Section 6.5.2. A MEX/Matlab imple-

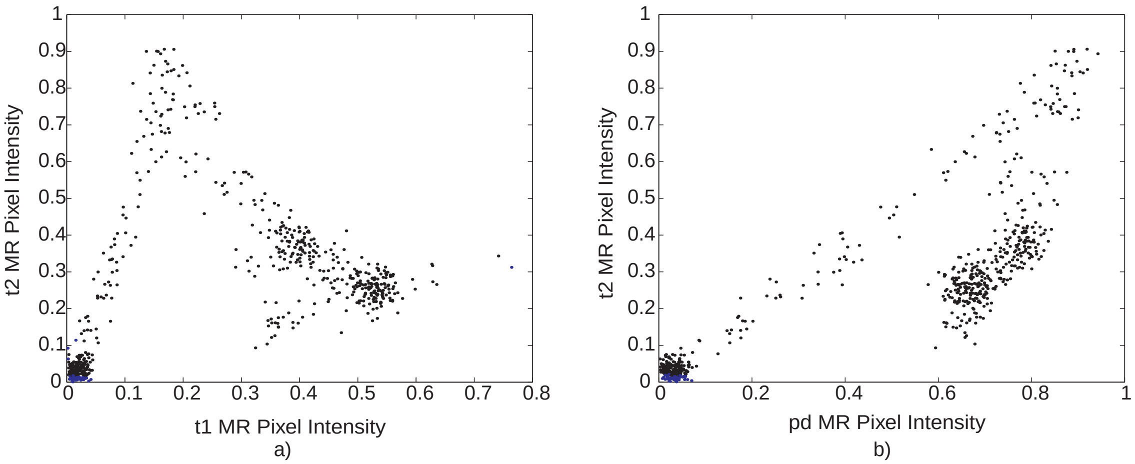
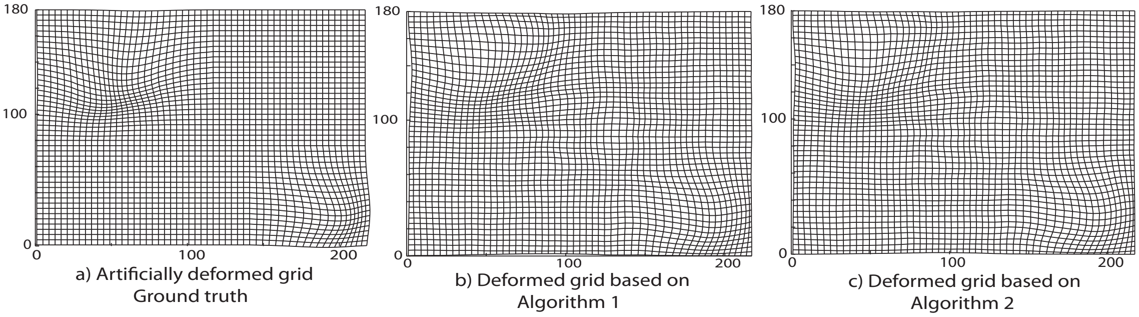


Figure 6.2: Pixel intensity samples from perfectly aligned image airs



a) Artificially deformed grid Ground truth

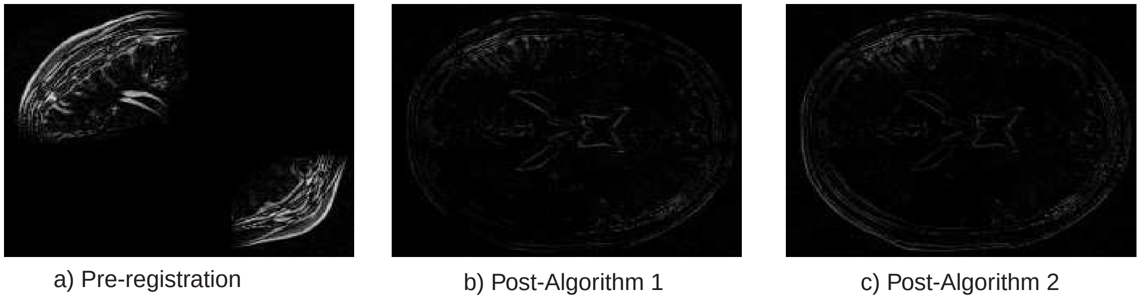
b) Deformed grid based on Algorithm 1

c) Deformed grid based on Algorithm 2

Figure 6.3: Grids deformed based on the original artificial warp and warps recovered using proposed algorithms.

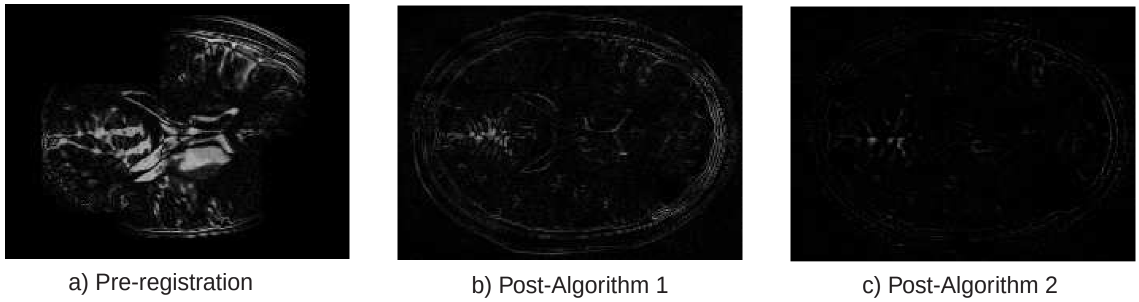
mentation running on 512MB Intel machine took about 44 seconds to complete the nonrigid alignment step. Figure 6.10 shows the two brains after nonrigid registration. Figure 6.3 illustrates the warp field obtained in the third step.

After inspecting the alignment of the edges in the images, e.g. the sulci and gyri, we conclude that both algorithms (EMST-based rigid body and level set nonrigid) yield promising results.



a) Pre-registration b) Post-Algorithm 1 c) Post-Algorithm 2

Figure 6.4: Difference images, i.e., absolute value of warped image - original image, before and after registration for Case 1, *t1-t2* registration.



a) Pre-registration b) Post-Algorithm 1 c) Post-Algorithm 2

Figure 6.5: Difference images, i.e., absolute value of warped image - original image, before and after registration for Case 2, *pd-t2* registration.

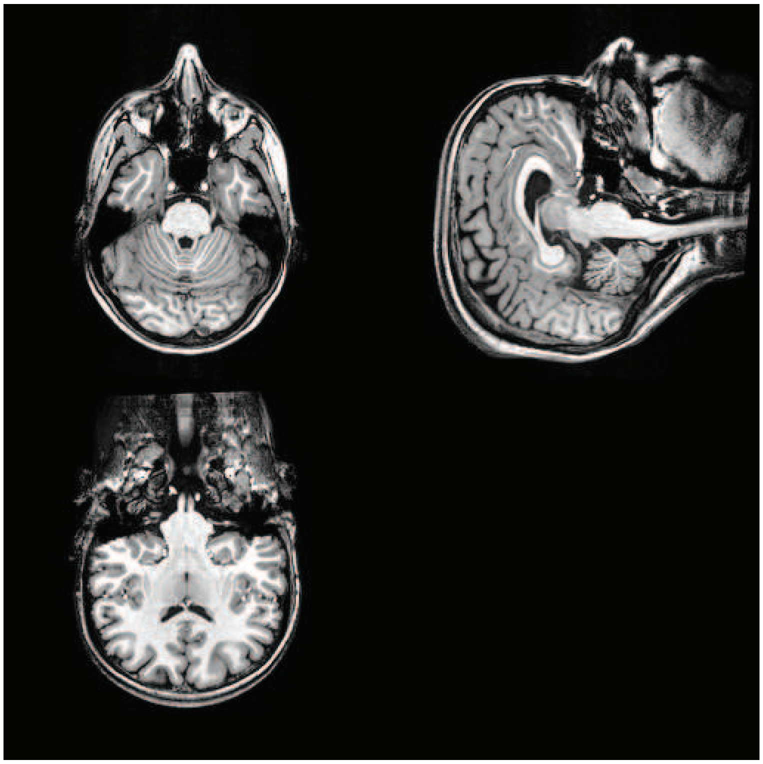


Figure 6.6: Subject 1: Transverse, sagittal and coronal views of the MR volume.

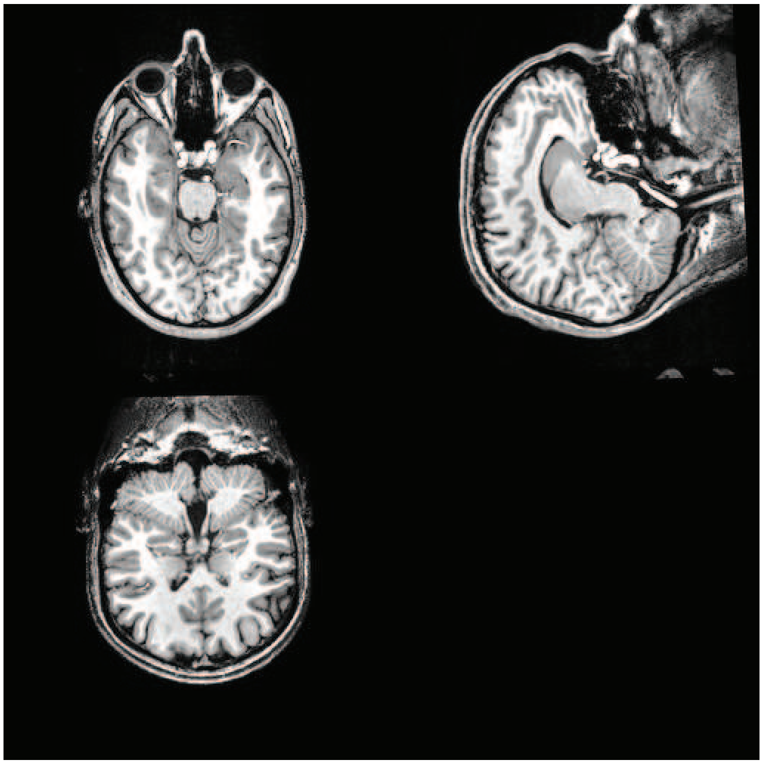


Figure 6.7: Subject 2: Transverse, sagittal and coronal views of the MR Volume

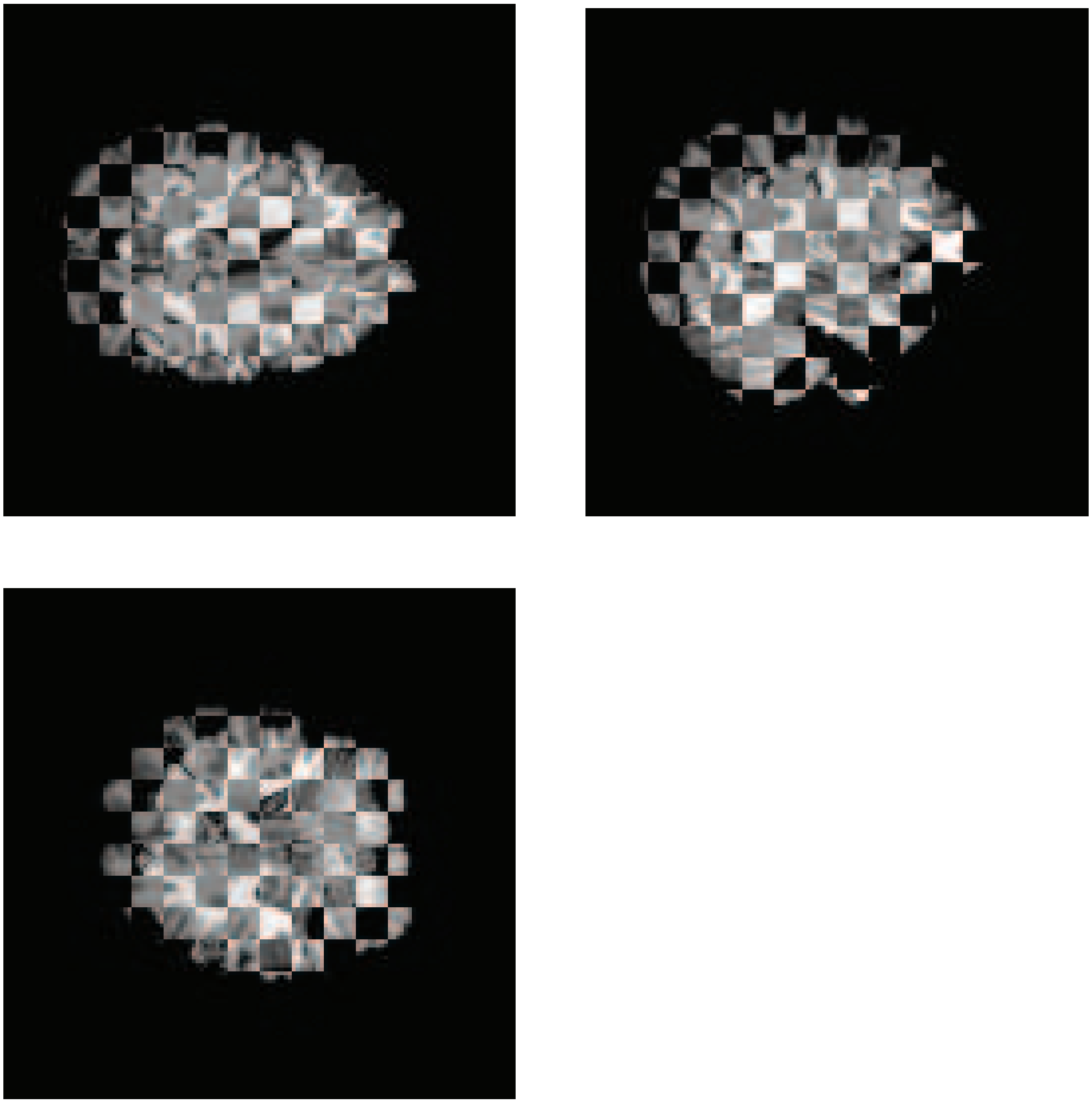


Figure 6.8: Checkerboard representations of the two subjects' brains at initial alignment: Transverse, sagittal and coronal views.

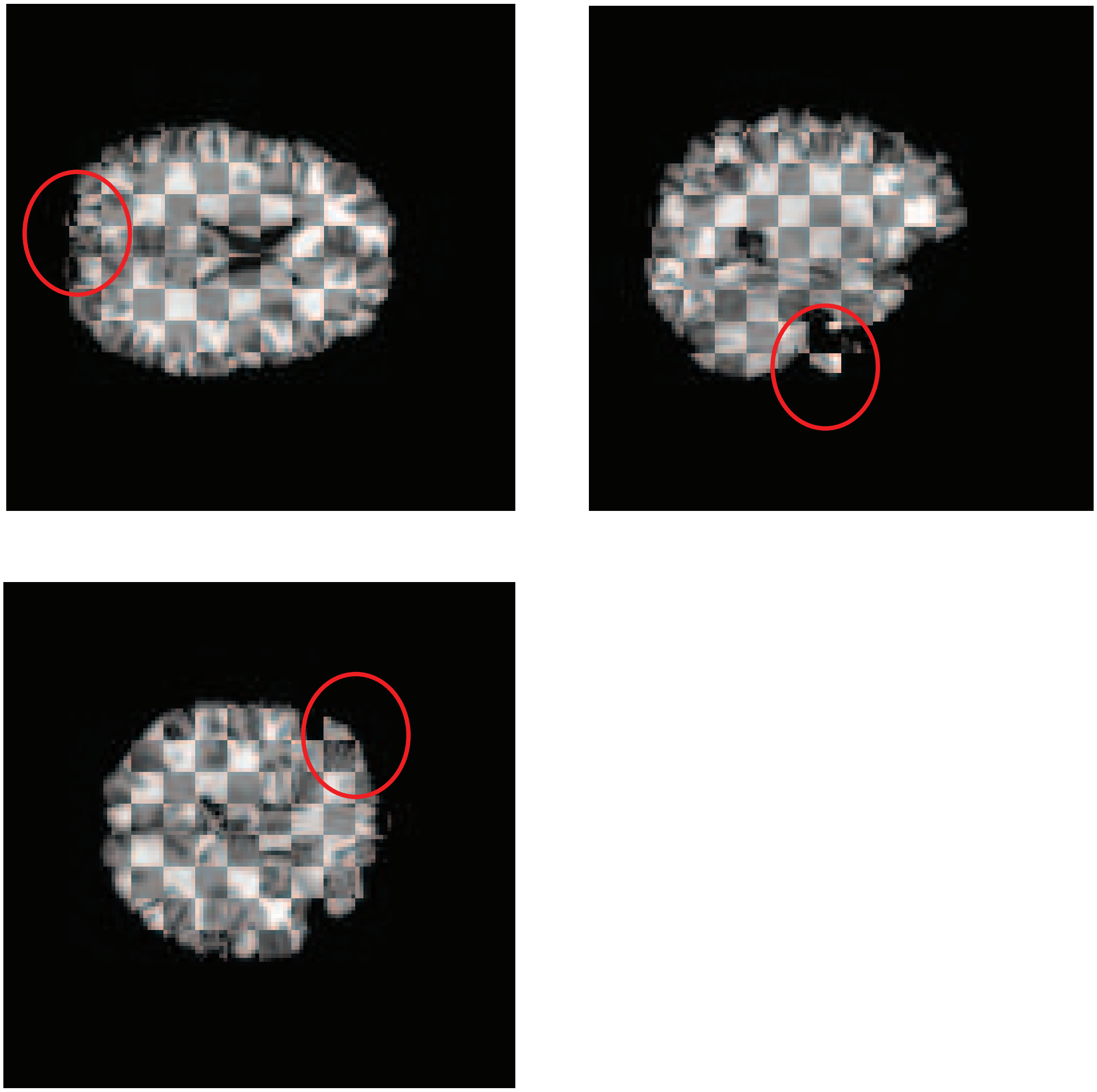


Figure 6.9: Checkerboard representations of the two subjects' brains after rigid alignment using the EMST-based algorithm described in Chapter 4: Transverse, sagittal and coronal views. Red circles indicate regions where local, nonrigid deformations are required to improve alignment.

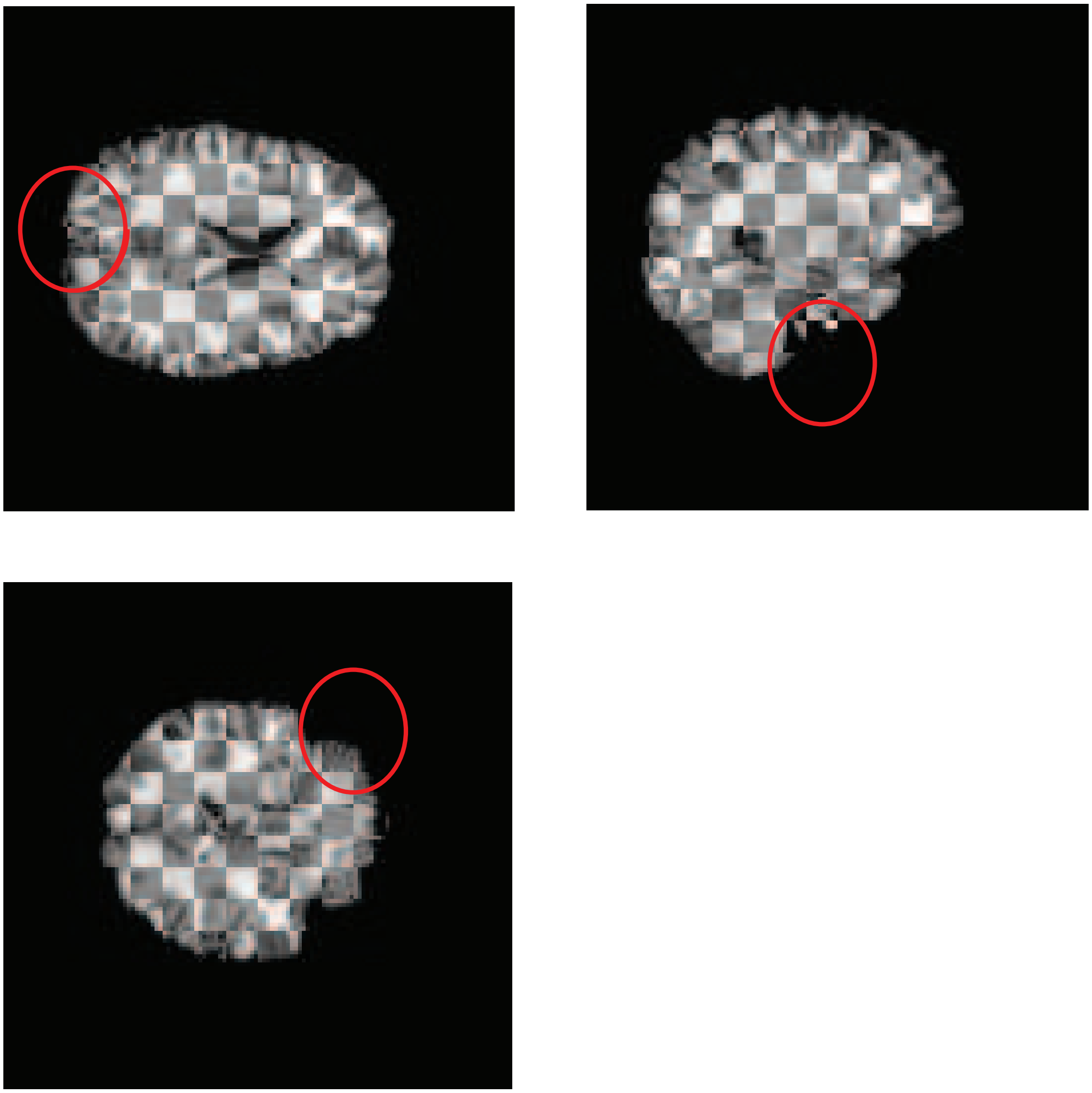


Figure 6.10: Checkerboard representations of the two subjects' brains after nonrigid alignment using Algorithm 1 described in Section 6.5.2: Transverse, sagittal and coronal views. The effect of the nonrigid alignment can easily be seen by comparing the alignment of the edges in regions within the red circles.

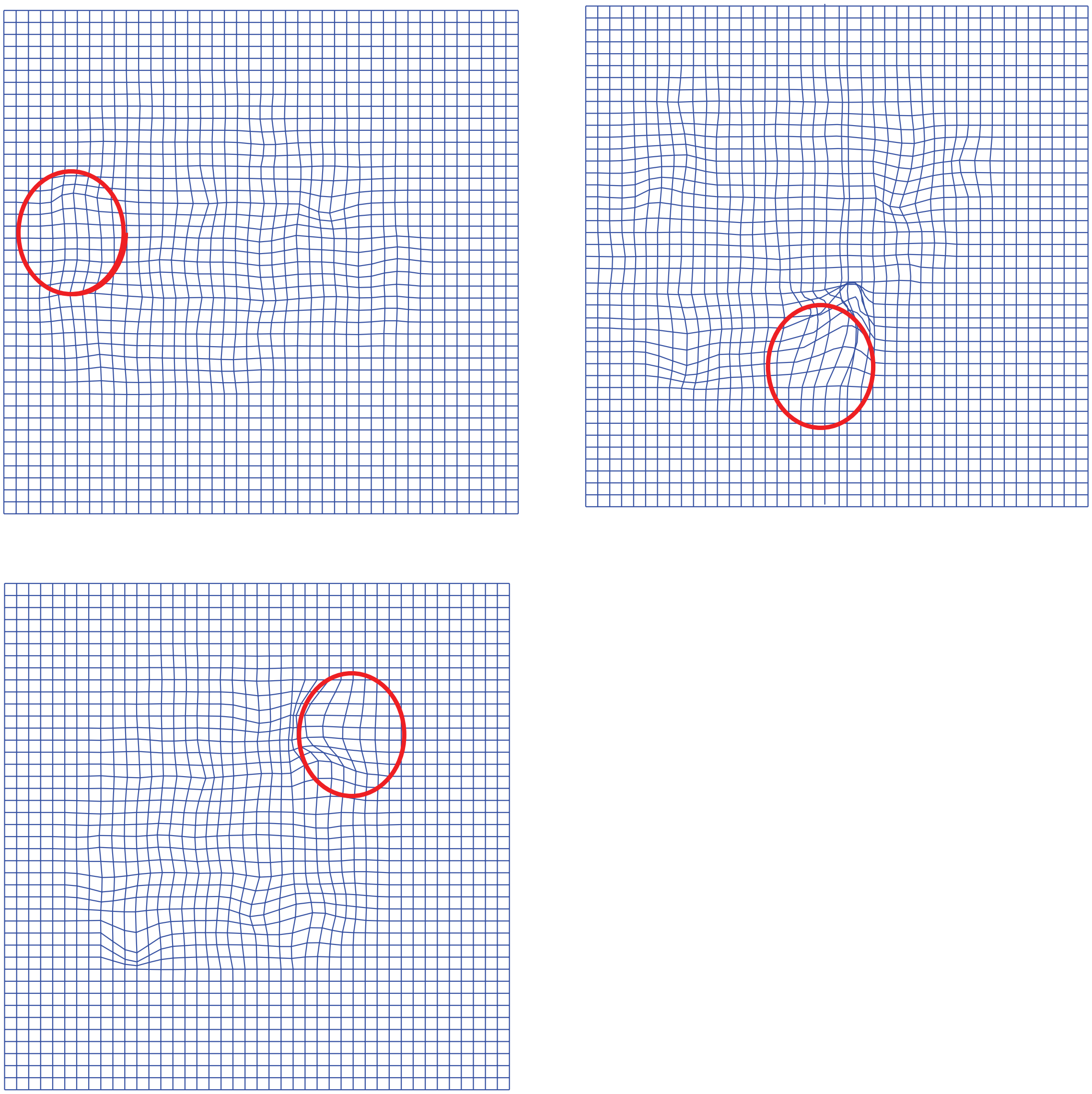


Figure 6.11: Warped grids from the nonrigid alignment step: Transverse, sagittal and coronal views. Red circles highlight regions where nonrigid alignment had an important role in lining up edges.

Chapter 7

Functional Registration of the Human Cerebral Cortex using fMRI

In this chapter, we depart from the theoretical aspects of entropy-based image registration and start the exploration of an application-driven problem: *the inter-subject functional registration of the human cerebral cortex*. To our knowledge, this is the first attempt at this problem.

The cerebral cortex is a large, continuous, 2-4 mm thick, folded sheet of tissue located right below the skull. An increasingly important part of today's neuroscientific research concerns the structural and functional organization of the cerebral cortex. However, a major challenge in the field is to set up correspondence between different subjects' brains, so that population-based conclusions can be drawn. This chapter is intended to serve as a summary of some preliminary work that addresses this challenge. We include a description of a proposed method that attempts to functionally align the cerebral cortices of different subjects using structural and functional MRI data gathered from multiple subjects during the viewing of a movie (Steven

Spielberg's "The Raiders of the Lost Ark"). Preliminary experiments indicate that the alignment results generalize well to other cognitive experiments, supporting the plausibility of functional normalization based on both structural and functional MRI.

7.1 Introduction

The cerebral cortex is a sheet-like, grey colored brain structure that is folded with deep involutions. These foldings create grooves (called sulci) and bumps (called gyri) that can be identified on the surface. The cerebral cortex is involved in many complex brain functions including memory, attention, perceptual awareness, language and consciousness. Based on scientific evidence (e.g. the columnar and laminar organization [58, 63, 71]), we believe that the functional organization of the cortex is intrinsically two-dimensional. In other words, most of the (functional or anatomical) features that distinguish different cortical areas can only be understood by viewing these regions on the folded manifold of the cerebral cortex. Moreover, in an unlabelled 3D volume, that is without the knowledge of functional and/or structural areas, there is no obvious way of explaining these phenomena. This characteristic of the cerebral cortex has inspired the design of tools that efficiently extract and represent this two-dimensional surface [30, 29, 2].

Setting up (functional or anatomical) correspondence across multiple subjects is a crucial precursor to most neuroscientific studies that try to understand how the brain functions and come up with useful models that can describe brain responses within a significantly large population. Most of today's studies that compare the functionalities of brains across multiple subjects rely on *anatomical normalization*, i.e., the registration of all subjects to an atlas (template) or average brain based on anatomical landmarks, such as major gyri and sulci, and/or high resolution structural MRI scans. The most common technique is the so-called Talairach normalization

[90], which is a 3D piecewise affine registration technique based on a small number of anatomical landmarks that include the posterior commissure (PC) and the anterior commissure (AC).

This chapter will heavily rely on the distinction between functional and structural neuroanatomy. Functional neuroanatomy reflects the organization and orientation of event-related and task-specific neural activity, whereas structural neuroanatomy refers to physical organization, such as the loci and orientations of sulci and gyri. There is significant evidence that suggests that functionally-defined regions are not consistently located relative to anatomical landmarks on the cerebral cortex. For example, the location of the visual motion area, MT, can vary across individuals by more than 2cm after Talairach normalization and can either be in the inferior temporal sulcus or the lateral occipital sulcus [92]. Moreover, the cortical area responsible for low-level visual processing, namely V1, can vary in size by as much as two-fold across different subjects' brains [70].

In this study, we investigate the use of patterns of neural activity evoked by cognitive and perceptual tasks as the basis for inter-subject registration of the functional cortical neuroanatomy. Our hope is that this research will lead to a general method for functional registration and the definition of a functional atlas of the human cerebral cortex. Our initial investigations have focused on employing the fMRI time series as indices of local functional response profile to perform non-linear registration on the convoluted two-dimensional manifold of the cortex surface. The following section contains a brief overview of functional MRI.

7.2 Functional MRI

Functional Magnetic Resonance Imaging (fMRI) produces videos (with today's technology, typically of 2-4 mm spatial resolution and 1-4 second temporal resolution) that represent the hemodynamic response related to neural activity in the brain. It is mainly based on the principle commonly known as Blood Oxygenation Level Dependent (BOLD) contrast. Neural activity leads to a temporary increase in the concentration of deoxygenated hemoglobin in the vicinity of the activity, which in turn intensifies the detectable BOLD signal due to the change in the blood magnetic susceptibility. This phenomenon has been extensively investigated. For further reading, the interested reader is referred to more dedicated works such as [57].

7.3 Pre-processing of the Data

We used tools from FreeSurfer[1] [30, 29] to obtain a tesselated representation of the cortical surface using high-resolution, T-1 weighted (structural) three-dimensional MRI volumes. This is a complex procedure (details of which can be found in [30]). In summary this procedure is broken into the following sub-tasks: Intensity-variances due to magnetic field inhomogeneities are corrected, "non-brain" regions are removed using a simple "skull-stripping" procedure, a segmentation procedure based on the geometric structure of the grey-white interface is performed, and a topologically correct segmentation of the white-matter is completed, resulting in a single filled volume for each cortical hemisphere. Finally, this volume is covered with a triangulation. This triangulation is then inflated and projected to a standard sphere using a procedure that preserves inter-node distances and the original topology [29]. Note that, at the end of this procedure, we obtain an irregular triangulation on a standard sphere, where each node's correspondence in the original 3D volume is known. To fix the irregularity of the triangular tessellation, the *mesh regularization* [2] tool in AFNI's SUMA[2] package was employed. At full spatial resolution, the regular mesh contained 144,002 nodes spaced 1mm apart. This regularization procedure uses a standard icosahedral

[1] FreeSurfer is a software package for the reconstruction of the cerebral cortical surface from structural MRI data, and the overlay of fMRI data on to the reconstructed surface

[2] AFNI is a software package for processing, analyzing, and displaying fMRI data

tessellation that is projected on to the standard sphere. Finally, after aligning the fMRI volumes with the structural MRI volume (using a multi-modal rigid-body registration algorithms, e.g. the EMST-based algorithm described in Chapter 4), each node of the regular mesh is assigned to a functional time-series, which is correlated with the neurological activity in the corresponding region in the brain.

7.4 Motivating Experiment

We have collected fMRI data from several subjects while they viewed an adventure movie (Steven Spielberg's "Raiders of the Lost Ark"), similar to the study by Hasson et al. [35]. Inter-subject correlations of voxel time activity curves have been calculated after Talairach normalization in the original 3D image space and after (structural) cortical surface normalization with FreeSurfer. The locations and strengths of inter-subject correlations were remarkably consistent with those reported in [35], in that with no spatial smoothing, the mean correlation for brain voxels was 0.05. Note that this correlation value is averaged over 144,002 nodes and is surprisingly high given the unconstrained nature of the experiment. There are also a small number of nodes that have a significantly large correlation value, e.g. grater than 0.3.

To obtain an initial estimate of how much more shared high spatial frequency signal might be recoverable with better methods of inter-subject alignment, we further analyzed data for two subjects on cortical surface models, dividing the data in halves. The cortical surfaces were structurally normalized based on cortical folding [30, 29]. Correlations between surface nodes with no spatial smoothing were similar to those for the Talairach normalized data in 3D space. To estimate the maximum between subject correlation that might be achievable with function-based alignment, for each node in subject A we found the nearby node (within a 3 cm radius) in subject B that was most strongly correlated with the subject A node. These "optimal node

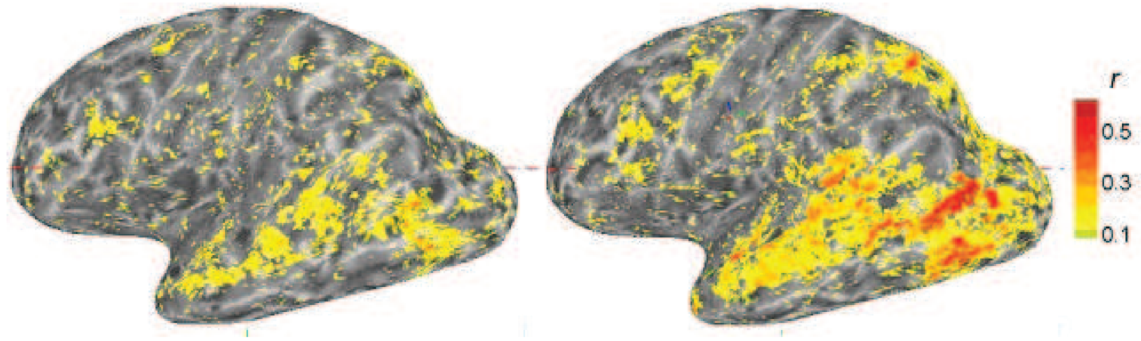


Figure 7.1: Correlations between responses in two subjects recorded during viewing a movie displayed on the inflated representation (FreeSurfer) of subject A. The correlations in both images are for node pairs in the second quarter of the data set (550 time points). The correlations in the left image were for cortical nodes with the same location after FreeSurfer-based structural registration. The correlations in the right image are for "optimal node matches" (obtained from the first quarter of the movie) that lay within 3 cm of each other.

matches" were determined from data acquired while subjects viewed the first quarter of the movie. This represents the upper boundary of shared variance that could be recovered with function-based alignment. For cross-validation of these "optimal node matches", we calculated the correlations of these "optimal node matches" in the second quarter of the movie. The mean correlation between "optimal node matches" in the second quarter of the data was twice the mean correlation from structurally-determined correspondences, i.e. after (structural) cortical surface normalization with FreeSurfer (see Figure 7.1). These results suggest that, with better methods of inter-subject alignment, we may be able to recover significantly more shared signal in an experimental paradigm.

7.5 Functional Registration using fMRI Data: Methodology

The basic idea of our proposed approach is to employ the whole fMRI time series data (that corresponds to some standard experiment, e.g. the viewing of an adventure movie) as a feature vector that represents the functionality of the corresponding point on the cortical surface. Anatomical alignment is used to initialize the algorithm. Thus, we view the functional registration algorithm as a fine-tuning of the anatomical alignment. In regions, where there is negligible activity detected by the fMRI scan, the algorithm will have no incentive to apply a warp, resulting in the preservation of the anatomical alignment. Moreover, instead of basing alignment on functionally-defined *areas*, whose location is usually defined as the center of mass or the local maximum response, the alignment is based on *patterns* of response as they are distributed spatially both within and across cortical areas [36]. In other words, the alignment is based on a complete correspondence code [4] that relates every cortical point in an individual's brain to a corresponding cortical point in the brains of other individuals. The proposed method is implemented on a standard two-dimensional representation (inflated and projected onto a standard spherical surface, as described in the previous section) of the cortical surface.

7.5.1 Correlation of the Time-series

As discussed in previous chapters, a crucial component of a registration algorithm is the alignment measure. This thesis has mainly investigated entropy-based alignment measures employed for different applications that can be classified as: *rigid, nonrigid and trained registration.* Inter-subject applications almost always require nonrigid registration. Here, we will employ a dense deformation approach, where each voxel/node is allowed to "move" independently. Note that this is slightly different than the approach detailed in Chapter 6, since an explicit penalty term is used to regularize the raw warp field, rather than Gaussian smoothing. Also, registration is performed on a spherical surface, not a Euclidean grid.

The dense deformation approach typically requires a local (point) alignment measure, the gradient of which determines the direction of the "move" (warp) of the corresponding point. When dealing with scalar images, all we have at a given point (voxel) is a scalar value. Thus, to compute the point-wise similarity between two images, we usually make use of the rest of the images (either locally, or as in Chapter 6, globally). In the case of fMRI, however, we have much richer information at each point: long time-series (of length 100-2000). We propose to make use of this information to compute a local (point-wise) alignment measure, the gradient of which can be used to drive the warp locally. Since computation time and memory are very valuable resources, and the data sets we are dealing with are extremely large, as an initial attempt, we investigated the correlation ρ between the two time-series as an alignment measure:

$$\rho(X,Y) = \frac{\mathbb{E}_{XY}((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y}, \tag{7.1}$$

where X and Y are random variables and μ and σ are the corresponding means and variances. In practice, we use sample mean estimates of μ and σ to compute an estimate of ρ. Note that, employing (7.1) as an alignment measure is equivalent to employing the conditional (Shannon) entropy $H(X|Y)$, under the assumption that $X = aY + N$, where a is an arbitrary, yet fixed scalar and N is i.i.d Gaussian noise. This is consistent with the General Linear Model (GLM) commonly used for fMRI time series analysis [99]. For a more detailed discussion of the link between correlation and entropy-based measures, the reader is referred to Section 4.1.2 of Viola's thesis [96].

7.5.2 Regularizing the Warp

Using correlation (7.1) as an alignment measure, one can compute the "optimal matches" for all nodes (through an exhaustive search), as presented in Section 7.4. The resulting correspondence can be considered as a *non-regularized functional align-*

ment of two subjects' cerebral cortices, since no regularization is imposed on the node matches[3].

As discussed in previous chapters, an algorithm that imposes no regularization on the warp field (e.g. exhaustive search), however, has the potential problem of overfitting to the data, and thus not generalizing well to new experiments. This is also the case in the functional registration of the cerebral cortex. For a registration result to be useful, we would like it to generalize well to new (test) experiments. That way, we can functionally register (normalize) subjects based on one standard experiment and use these results with other experiments. The hope is that this (*functional normalization*) procedure will yield improved results for studies that investigate the functionality of the human cerebral cortex within a group of subjects.

To illustrate the effect and importance of regularization (or in this case, the lack of it), we conducted a simple experiment. An exhaustive search algorithm determined the "optimal node matches" for two subjects. Note that, each optimal match has a corresponding match score: the correlation value between the time-series in the two subjects. Next, the algorithm sorted these matches w.r.t their match scores in descending order. Thus, the first optimal match in this list is a pair of node indices (from each subject), that indicate a functional correspondence (determined by the exhaustive search), and has the highest correlation value. This can be viewed as the point of best functional correspondence between the two subjects. Next, for the generalization test, we used these optimal matches on a new experiment (visual category [36]) data set. Figure 7.2 shows the variation of the average, per node correlation between the left hemispheres of two subjects with respect to the number of optimal matches (determined using the "movie" experiment) used from the sorted list. The shape of this curve supports the idea that node matches with high correlation values generalize to a new experiment, whereas weaker correspondences do not.

[3] Strictly speaking, there was minimal regularization: the exhaustive search was conducted over all nodes (in the second subject) within a 3 cm distance of the seed node.

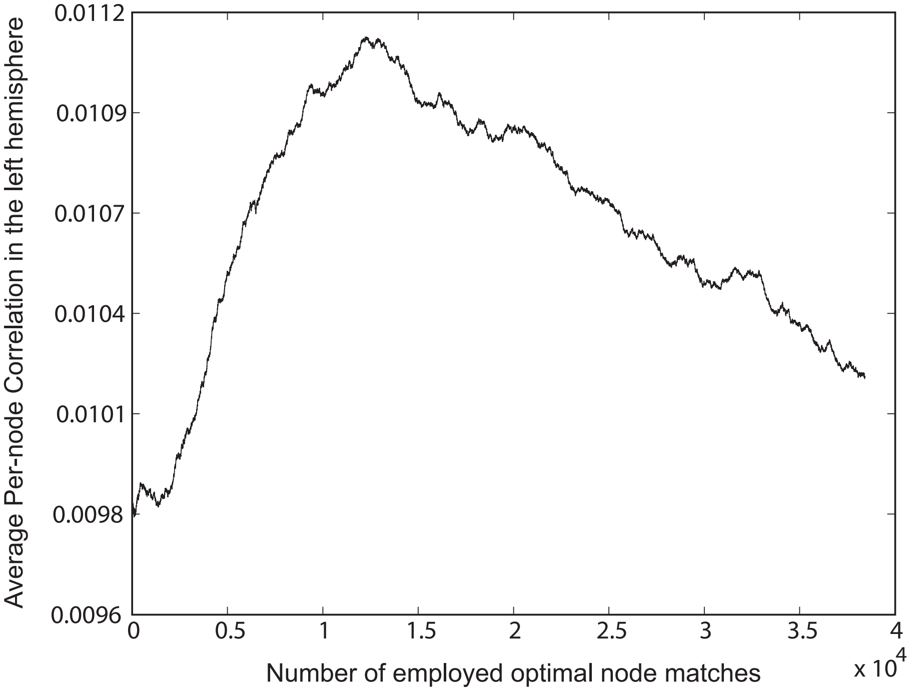


Figure 7.2: The variation in the average per-node correlation between the two subjects for the test ("category recognition") experiment with respect to the number of optimal matches (from the "movie part 1") used to set up functional correspondence. Zero matches is equivalent to the anatomical alignment.

There are many approaches to regularize a non-linear spatial warp. Section 6.1 contains a brief overview of these approaches. The main goal is to avoid over-fitting by incorporating our expectations about the warp. This is usually achieved by penalizing unexpected warps. Typically, *invertibility* and *smoothness* are the two main characteristics imposed on a spatial warp. In the triangulated (mesh-like) representation of the cortical surface, smoothness is related to the preservation of inter-node distances. Invertibility, on the other hand, can be achieved by avoiding "foldings" of the mesh.

To avoid over-fitting in the context of functional registration of the human cerebral cortex, we propose to use one of the weakest constraints on the warp, specifically avoiding the folding of the mesh. This choice is partly due to the lack of strong scientific evidence that would justify any other regularization. For instance, as we discussed in Section 7.1, the areas of some well-defined functional regions can vary significantly across individuals. This suggests that imposing the preservation of inter-node distances may be too strict for this specific application. However, Figure 7.2 indicates that employing all optimal node matches is not the optimum functional alignment. Moreover, the mesh obtained by applying all the optimal matches contains many foldings and the resulting warp is not invertible. Similar to [29], as a method of mild regularization, we investigated the employment of an aerial distortion penalty term in our alignment measure. This term effectively prevents folding in the warped mesh, but puts no constraints on inter-node distances.

7.6 Implementation

The cortical surface is represented with a regularized triangulation, which is stored as a list of mesh-nodes. For each subject i, each mesh-node v contains a spatial position $\mathbf{x}_v^i$, experimental time-series $\mathbf{t}_v^i$, a list of neighboring nodes N_v, and belongs to a list of mesh triangles T_v. In the regularized mesh, all inter-neighbor distances are the same, and all triangles have the same area A_0.

The functional registration algorithm modifies the time-series and spatial positions of the floating subject mesh-nodes only. This is stored as a warp-field[4], which can be added to the original spatial positions to interpolate the new time-series.

The algorithm attempts to maximize $E = E_c - \lambda E_a$, where E_c is the total node-wise inter-subject correlations (i.e., the alignment measure), E_a is the areal penalty term (i.e., the regularization) and λ is a scalar weight that determines the influence of the regularization term. Each node is allowed to move independently and the optimization is done using gradient-ascent.

Let $\mathbf{s}_v^i \triangleq \mathbf{t}_v^i - \bar{\mathbf{t}}_v^i$, where $\bar{\mathbf{t}}_v^i$ is the mean value of the time-series. Then the alignment measure (between subjects i and j) can be written as:

$$E_c(i,j) = \sum_v \hat{\rho}(\mathbf{t}_v^i, \mathbf{t}_v^j) = \sum_v \frac{\mathbf{s}_v^i \cdot \mathbf{s}_v^j}{|\mathbf{s}_v^i||\mathbf{s}_v^j|}, \tag{7.2}$$

where $\hat{\rho}$ is the sample correlation and $|.|$ denotes the magnitude of a vector. The gradient of $E_c(i,j)$ with respect to the spatial position $\mathbf{x}_v^j$ is:

$$\frac{\partial E_c(i,j)}{\partial \mathbf{x}_v^j} = \hat{\rho}(\mathbf{t}_v^i, \frac{\partial \mathbf{t}_v^j}{\partial \mathbf{x}_v^j}). \tag{7.3}$$

Let A_{uvw} denote the oriented area of the mesh triangle Δ that consists of mesh nodes u, v and w ($u > v > w$); $\mathbf{x}_{uv}^j = \mathbf{x}_v^j - \mathbf{x}_u^j$ and $\mathbf{n}_u^j = \frac{\mathbf{x}_u^j}{|\mathbf{x}_u^j|}$ is the surface normal at node u. Define:

$$A_\Delta^j = A_{uvw}^j = \mathbf{x}_{uv}^j \times \mathbf{x}_{uw}^j \cdot \frac{\mathbf{n}_u^j}{2}. \tag{7.4}$$

Let A_Δ^{j0} denote the oriented area of triangle Δ in subject j's regularized mesh. Similar

[4] The warp field is in spherical coordinates, since the mesh-nodes are only allowed to move on the spherical surface, conforming to the two-dimensional topology of the cortex

to [29], we define the areal penalty term as:

$$\begin{aligned} E_a &= \frac{1}{3}\sum_{v}\sum_{\Delta\in T_v}(A^j_\Delta - A^{j0}_\Delta)^2 I(A^j_\Delta, A^{j0}_\Delta), \qquad (7.5)\\ I(A^j_\Delta, A^{j0}_\Delta) &= \begin{cases} 1 & \text{if } A^j_\Delta A^{j0}_\Delta < 0, \\ 0 & \text{else.} \end{cases} \end{aligned}$$

In other words, if a mesh triangle is folded, the penalty is proportional to the difference between the current area and original area. Otherwise, it is zero. The gradient of E_a with respect to the spatial position $\mathbf{x}^j_u$ is:

$$\frac{\partial E_a}{\partial \mathbf{x}^j_v} = 2\sum_{\Delta\in T_v}(A^j_\Delta - A^{j0}_\Delta)I(A^j_\Delta, A^{j0}_\Delta)\frac{\partial A^j_\Delta}{\partial \mathbf{x}^j_v} \qquad (7.6)$$

Based on a gradient ascent framework, the algorithm can be summarized with the following update equation:

$$\mathbf{x}^j_v(t) \;=\; \mathbf{x}^j_v(t-1) + \varsigma(t)(\frac{\partial E_c(i,j)}{\partial \mathbf{x}^j_v} - \lambda\frac{\partial E_a}{\partial \mathbf{x}^j_v})|_{t-1},$$

where $\varsigma(t)$ is a step size.

Note that gradient-ascent finds the local optimum, and thus to find the global optimum, it is important to have a good guess for the initial values $\mathbf{x}^j_v(0)$. In our implementation, we employ exhaustive search results (i.e., the optimal node matches) to initialize the warp. The search is conducted within a 3 cm radius of the anatomical correspondence. We, then, compute a raw warp field using the node matches that have a score (correlation value) greater than some threshold value (typically $0.1-0.3$). Finally, this warp field is smoothed with an approximately Gaussian filter using AFNI's surfsmooth tool. The smoothed warp field is used to initialize the iterative gradient ascent.

7.7 Empirical Results

In this section, we include preliminary results obtained from four subject pairs: rb-kd, cb-kl, dm-mh and ph-se. The second subjects in each pair were functionally aligned to the respective first subjects using the procedure described in previous sections. The fMRI data gathered while subjects were viewing the first half of Steven Spielberg's "Raiders of the Lost Ark" (movie P1) was used for alignment. Generalization tests were performed using the second half of the viewing (movie P2) and a visual category (vis. cat.) experiment. Table 7.1 lists the correlation values between the anatomically and functionally aligned pairs in the whole brains (both hemispheres: lh, rh). Table 7.2 contains correlation scores for an anatomically defined region of interest (specifically the ventral temporal cortex, which is active in face and object recognition). The generalization results seem to be fairly consistent and indicate that using fMRI data from movies part 1, we can improve the correlation values for other test experiments by $5 - 20\%$. Our C++ implementation took an average run-time of 20-25 minutes for the functional registration of two subjects.

Subjects/Hemisphere	**Movie P1**		**Movie P2**		**vis. cat.**	
	anat.	func.	anat.	func.	anat.	func.
rb-kd/lh	0.0419	0.0790	0.0489	0.0557	0.0101	0.01098
rb-kd/rh	0.047	0.084	0.049	0.055	0.0118	0.0125
cb-kl/lh	0.053	0.0897	0.0461	0.0569	0.0156	0.0174
cb-kl/rh	0.0582	0.0932	0.0513	0.0586	0.0159	0.0173
dm-mh/lh	0.0494	0.0807	0.0490	0.0552	0.0128	0.0144
dm-mh/rh	0.0565	0.0876	0.056	0.0628	0.0158	0.0175
ph-se/lh	0.035	0.0703	0.0395	0.043	0.0048	0.0052
ph-se/rh	0.0417	0.0786	0.0478	0.053	0.0069	0.0075
mean	0.04784	0.08288	0.04846	0.05506	0.011771	0.01284

Table 7.1: Per node, averaged (within whole hemispheres) correlation values between fMRI time-series of subject pairs. anat.: FreeSurfer based anatomical alignment; func.: Functional alignment. For a detailed description, see text.

Figure 7.3 shows the node correlations between subjects rb and kd before functional registration (i.e., at anatomical registration) and after functional registration

Subjects/Hemisphere	**Movie P1**		**Movie P2**		**vis. cat.**	
	anat.	func.	anat.	func.	anat.	func.
rb-kd/lh	0.083	0.1149	0.076	0.0872	0.022	0.244
rb-kd/rh	0.0953	0.124	0.0722	0.0792	0.0252	0.0267
cb-kl/lh	0.1051	0.1464	0.0917	0.1156	0.0399	0.0457
cb-kl/rh	0.111	0.1443	0.0968	0.1095	0.0428	0.0452
dm-mh/lh	0.0782	0.1077	0.0715	0.0812	0.0402	0.0444
dm-mh/rh	0.0971	0.1273	0.0848	0.0957	0.0437	0.0484
ph-se/lh	0.0557	0.0844	0.0515	0.0572	0.0134	0.0148
ph-se/rh	0.0621	0.0918	0.0655	0.0729	0.0151	0.0166
mean	0.0859	0.1176	0.0762	0.0873	0.0303	0.0333

Table 7.2: Per node, averaged (within the Ventral Temporal Cortex) correlation values between fMRI time-series of subject pairs. anat.: FreeSurfer based anatomical alignment; func.: Functional alignment. For a detailed description, see text.

for the three experiments: movie part 1 (the experiment used to functionally register the data sets), movie part 2 and the visual category experiment.

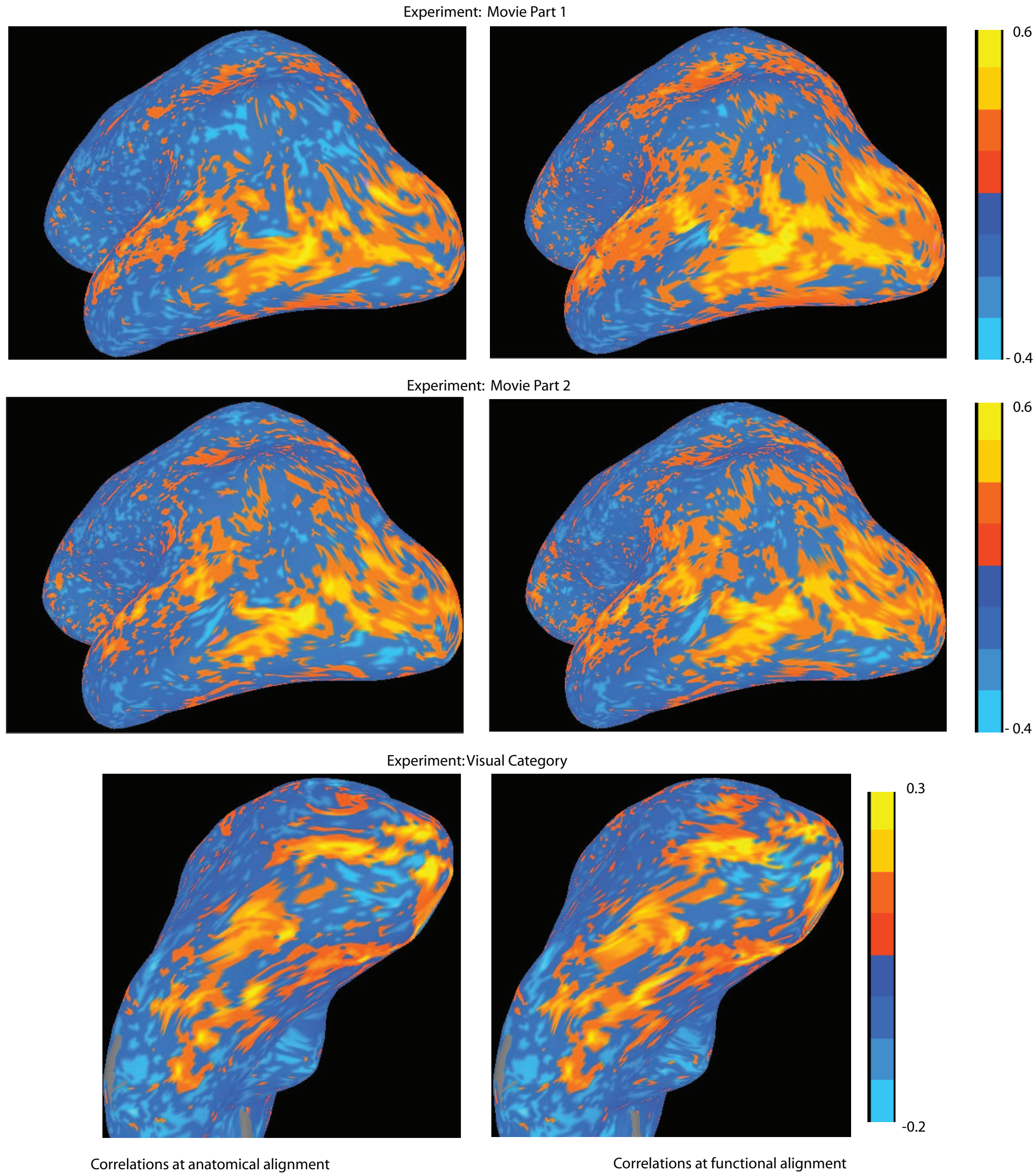

Experiment: Movie Part 1
0.6
- 0.4
Experiment: Movie Part 2
0.6
- 0.4
Experiment:Visual Category
0.3
-0.2
Correlations at anatomical alignment
Correlations at functional alignment

Figure 7.3: Correlations between responses in two subjects recorded during viewing a movie displayed on the inflated representation (FreeSurfer) of subject rb. The (color-coded) values are correlations between rb and kd's corresponding time-series for three experiments: movie part 1 and 2, and visual category. The correlations in the left images are for cortical nodes with the same location after FreeSurfer-based anatomical registration. The correlations in the right images are after functional alignment (based on the first half of the movie experiment). The first two rows of images is a sagittal view of the whole left hemisphere. The last row is a ventral view of the VT cortex.

Chapter 8

Conclusions

In this thesis, we investigated algorithms to spatially align two, three or four dimensional digital images. This problem is particularly difficult when the images are obtained through different sensor types (multi-modal registration) and/or when complex nonlinear geometric transformations are required to relate the images, e.g. when registering different human brains (inter-subject registration).

In Chapter 2, we provided theoretical motivation for the employment of information-theoretic measures for multi-modal image registration. Chapter 3 focused on the entropy estimation problem and included a novel comparison of different entropy estimators from the perspective of image registration. This comparison provided valuable insight on how these techniques weight data which lead to predictions of likely performance when applied to image registration. These interpretations were confirmed by simulation results.

The comparison of entropy estimators and a thorough analysis of the differentiability problem of the entropic graph based estimator lead to a novel Rényi entropy-based registration framework detailed in Chapter 4. This framework, which is the main contribution of the first half of the thesis, yields fast and accurate multi-modal rigid registration algorithms.

In certain real-world applications, previously registered image pairs are available to the algorithm. In a multi-modal setting, these pre-registered images contain valuable information about cross-modality relationship. In Chapter 5, we included an overview of this problem and proposed a method for incorporating prior information about the modality relationship from pre-aligned image pairs into the entropic graph-based registration framework. Experimental results suggest that this improves the capture range of the alignment measure and makes it more robust against bad initial alignment.

In Chapter 6, we presented a fast (linear time in the number of pixels) entropy-based nonrigid image registration algorithm. The proposed method employs a "level set entropy" similarity measure, which can be derived using a maximum likelihood approach. The level set entropy formulation has two major advantages: since it is a one-dimensional entropy, fast entropy estimators, e.g. histogram-based methods, that suffer in high dimensional spaces can easily be used. Moreover, it is easier to make stronger assumptions within each level set, which allows the use of parametric models that yield faster and/or more accurate registration algorithms. One example, suitable for applications where the cross-modality relationship is one-to-one, is a Gaussian density model. We also demonstrated that the method of smoothing the gradient field is equivalent to employing a gradient-descent optimization strategy with a particular class of smooth transformations. The relationship with re-gridding and invertibility conditions were briefly discussed.

Finally, in Chapter 7 we included a discussion of a preliminary investigation of an interesting scientific problem: the functional alignment of the human cerebral cortex. We explored a simple algorithm to functionally register brains based on the structural and functional MRI data gathered while subjects were viewing an adventure movie. Experimental validation performed to data is promising, since it indicates that the proposed tool produces results that generalize to other cognitive experiments.

8.1 Future Research

In the Rényi Entropy-based registration framework, we mainly focused on rigid-body problems. Extension of these ideas to richer transformations, e.g. affine and spline based models, is crucial for many real-world applications and should be pursued. Note that we included a brief discussion of this issue is Section 4.1.

A major challenge in the current literature is to perform population registration on large collections of data sets. Currently available tools typically attack this problem by pre-selecting a reference data set (template) and registering in a pairwise fashion. The computational cost and potential inaccuracy of this approach can be eliminated by performing a simultaneous registration on the whole population. Moreover, population registration could help determine sub-groups of data sets, e.g. normal and abnormal brains, and make inferences by observing the variability within and between these sub-groups. The methods, we have explored in this thesis, such as descent-based registration using entropic graphs, seem to have the desirable theoretical properties and computational speed for achieving population registration. An immediate next step would be to investigate this open problem that may lead to a significant contribution.

In Chapter 6, we proposed a level set entropy measure as a similarity metric for nonrigid registration. We employed a simple technique that smoothed the raw gradient field to iteratively warp the image. Other regularization techniques can also be investigated with the level set entropy measure.

The functional alignment of the human cerebral cortex is a very promising project that is still at a preliminary stage. More experiments need to be conducted to validate and evaluate the proposed tool. Moreover, other types of data, e.g. diffusion tensor images, might be employed to achieve functional alignment. We consider this area as an important direction for future research.

Appendix A

Limit of Rényi Entropy

Here, we include a simple proof for the statement that Shannon's entropy is the limit of the Rényi entropy:

$$\begin{aligned}
\lim_{\alpha \to 1} H_\alpha(X) &= \lim_{\alpha \to 1} \frac{1}{1-\alpha} \log(\sum_x p_X(x)^\alpha) && \text{(A.1)} \\
&= -\frac{\sum_x p_X(x)^\alpha \log p_X(x)}{\sum_x p_x(x)^\alpha} && \text{(A.2)} \\
&= -\sum_x p_X(x) \log p_X(x) && \text{(A.3)} \\
&= H(X). && \text{(A.4)}
\end{aligned}$$

The second equation is the application of L'Hopital's rule. Note that, for the continuous case, we can replace the sums with integrals.

Appendix B

Histogram-based Entropy Estimator

A histogram-based entropy estimator employs a finite sum to approximate the expectation in the entropy formula. Let $\mathbb{Q}$ be a countable subset of $\mathbb{R}^d$ that includes the origin and is closed under addition and subtraction, $q(\mathbf{x}) \in \mathbb{Q}$ denote the quantized ("binned") value of $\mathbf{x} \in \mathcal{X}$, $K(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}$ be a symmetric density, $h(k; \mathcal{X})$ denote the number of samples $\mathbf{x} \in \mathcal{X}$ that satisfy $q(\mathbf{x}) = k$ and the total number of samples be N. Define the discrete kernel:

$$\bar{K}(\mathbf{z}) \triangleq \frac{K(\mathbf{z})}{\sum_{\mathbf{m} \in \mathbb{Q}} K(\mathbf{m})}, \forall \mathbf{z} \in \mathbb{Q}.$$

A Parzen-window estimate (3.4) of the p.m.f. of the quantized random variable $q(X)$ is:

$$\begin{aligned} \hat{p}_H(\mathbf{z}; \mathcal{X}) &= \frac{1}{N} \sum_{i=1}^{N} \bar{K}(\mathbf{z} - q(\mathbf{x}_i)) \\ &= \frac{1}{N} \sum_{m \in \mathbb{Q}} h(\mathbf{m}; \mathcal{X}) \bar{K}(\mathbf{z} - \mathbf{m}) = \frac{1}{N} \sum_{m \in \mathbb{Q}} h(\mathbf{z} - \mathbf{m}; \mathcal{X}) \bar{K}(\mathbf{m}). \end{aligned}$$

A histogram-based estimate of the α information potential using a finite sum expectation is:

$$\begin{aligned}
\hat{V}_H(\mathcal{X},\alpha) = E_{\mathbb{Q}}(\hat{p}_H^{\alpha-1}(.;\mathcal{X})) &\triangleq \sum_{\mathbf{m}\in\mathbb{Q}} \hat{p}_H(\mathbf{m})\hat{p}_H^{\alpha-1}(\mathbf{m}) \\
&= \sum_{\mathbf{m}\in\mathbb{Q}} \frac{1}{N} \sum_{\mathbf{n}\in\mathbb{Q}} h(\mathbf{m}-\mathbf{n})\bar{K}(\mathbf{n})\hat{p}_H^{\alpha-1}(\mathbf{m}) \\
&= \frac{1}{N} \sum_{\mathbf{m}'\in\mathbb{Q}} \sum_{\mathbf{n}\in\mathbb{Q}} h(\mathbf{m}')\bar{K}(\mathbf{n})\hat{p}_H^{\alpha-1}(\mathbf{m}'+\mathbf{n}) \\
&= \frac{1}{N} \sum_{\mathbf{m}'\in\mathbb{Q}} \sum_{\mathbf{n}'\in\mathbb{Q}} h(\mathbf{m}')\bar{K}(\mathbf{n}')\hat{p}_H^{\alpha-1}(\mathbf{m}'-\mathbf{n}') \\
&= \sum_{\mathbf{m}\in\mathbb{Q}} \bar{K}(\mathbf{m}) \frac{1}{N} \sum_{\mathbf{n}\in\mathbb{Q}} h(\mathbf{n})\hat{p}_H^{\alpha-1}(\mathbf{n}-\mathbf{m}) \\
&= \sum_{\mathbf{m}\in\mathbb{Q}} \bar{K}(\mathbf{m}) E_{q(\mathcal{X})}(\hat{p}_H^{\alpha-1}(.-\mathbf{m};\mathcal{X})) \\
&= \sum_{\mathbf{m}\in\mathbb{Q}} \bar{K}(\mathbf{m}) E_{q(\mathcal{X})-\mathbf{m}}(\hat{p}_H^{\alpha-1}(.;\mathcal{X})),
\end{aligned}$$

where $E_{\mathcal{X}}$ is the sample mean on $\mathcal{X}$, $q(\mathcal{X}) + m \triangleq \{q(\mathbf{x}_i) + m : \mathbf{x}_i \in \mathcal{X}\}$, and $\hat{p}(\cdot)$ and $h(\cdot)$ are short-hand notations for $\hat{p}(.;\mathcal{X})$ and $h(.;\mathcal{X})$, respectively. Note that if $K(m) = 1$ iff $m = 0$ and zero otherwise, then $\hat{V}_H(\mathcal{X},\alpha) = \hat{V}_M(q(\mathcal{X}),\alpha)$.

Appendix C

Families of Graphs

In this section, we provide brief definitions of three different families of graphs. The minimal graphs (3.8) that correspond to these families have continuous and quasi-additive weights [40] and thus can be used to estimate the underlying entropy, as discussed in Section 3.3.3. See Figure C.1 for examples.

- **Spanning Tree:** A spanning tree of a vertex set V is a connected, acyclic, undirected graph that spans all vertices in V. Without the connectivity requirement, the graph is called a *spanning forrest.* The graph that has the minimum total weight amongst all spanning trees is called a *Minimum Spanning Tree* (MST). If the edge weights are defined as Euclidean distances, then the MST is a Euclidean MST (EMST).

- **Hamiltonian Cycle:** A Hamiltonian cycle of a vertex set V is an undirected graph that visits each vertex only once and also returns to the starting vertex. If it doesn't return to the starting vertex, it is called a Hamiltonian path. Also, a graph that contains a Hamiltonian cycle is called a *Hamiltonian graph.* The problem of searching for the Hamiltonian graph with minimum total weight is called the travelling salesman problem (TSP).

- **k-Neighbor Graph:** We define a k-Neighbor graph as a directed graph, where each vertex is the tail of k edges directed to *other* vertices. The corresponding minimal graph is called the k-nearest neighbor graph (kNN), or simply nearest neighbor graph if $k = 1$.

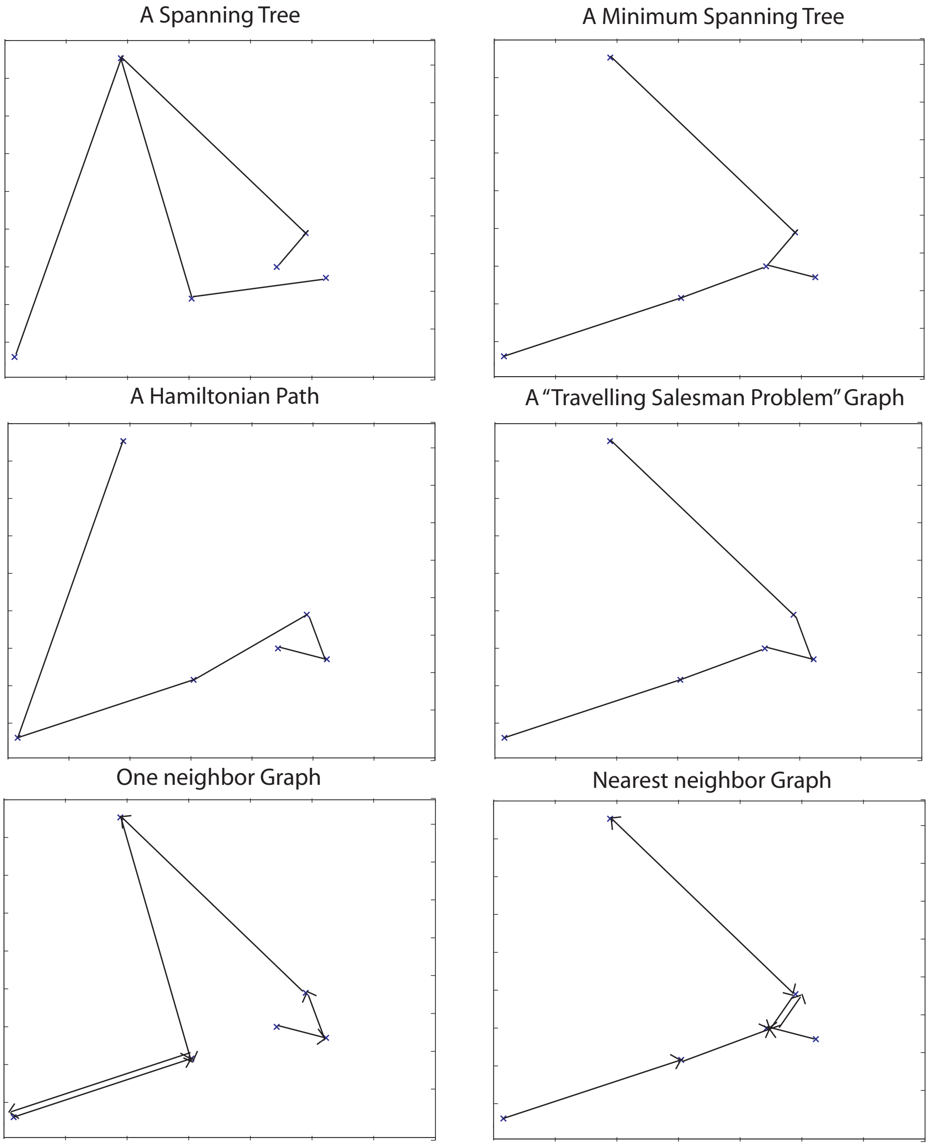


Figure C.1: Examples for various types of graphs and corresponding (Euclidean) minimal graphs.

Appendix D

Decomposing the Adjacency Matrix

An adjacency matrix $A(G)$ contains the topology information of a graph G. The (i, j)th entry $A(G)(i, j)$ is the number of edges connecting vertices i and j. Note if G is an undirected graph, $A(G)$ is symmetric. The following result allows us to derive (3.11) and hence recognize the entropic graph method as a special case of the plug-in estimator.

Theorem D.0.1. *Let G be a graph that contains at most one cycle in each of its connected components. There exists a matrix L such that $L + L^T = A(G)$ and in each of its rows is either a standard basis or a zero vector.*

Proof. First, let's assume G is a connected graph. Let's use mathematical induction to prove the existence of L.

1. For 2 vertices: If G contains no cycles, define

$$L = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

If G contains a cycle, define:

$$L = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

It is easy to see that these L's satisfy the necessary conditions.

2. Assume the result holds for $N-1$ vertices with L_{N-1}.

3. Let G_N be a graph with N vertices.

   - If G_N contains no leaf vertices, i.e., is one circular path, define L_N using the following algorithm (**A1**): Starting from the N'th vertex, traverse the circular path in the direction of the neighbor with the largest index. Let the i'th row of L_N be the j'th standard basis, where j is the vertex that follows the i'th vertex in the path.

   - If G_N contains at least one leaf vertex, define L_N using the following algorithm (**A2**): Let i be the index of the leaf vertex connected with the longest edge (if there are more than one of these, pick the one with the largest index amongst the candidates). Let $G_{N-1}(i)$ be the graph generated by pruning the i'th vertex (and its edge) in G_N. Let $L_{N-1}(i)$ be the matrix computed from $G_{N-1}(i)$ using **A1**, **A2** and/or Step 1. Define L_N by inserting an i'th row (the j'th standard basis, where j is the vertex connected to the i'th vertex in G_N) and i'th column (zero vector) to $L_{N-1}(i)$.

It can be seen that this L_N satisfies the desired properties.

If G is not a connected matrix, we can define a block-diagonal matrix L, where each of the diagonal blocks L_i correspond to the connected subgraph G_i and satisfy the desired properties. □

Corollary D.0.2. *If G is a spanning forest, a nearest-neighbor graph, a Hamiltonian cycle or a TSP, there exists a matrix L such that $L + L^T = A(G)$ and each of its rows is either a standard basis or a zero vector.*

Proof trivially follows from Theorem D.0.1.

Appendix E

Differentiability of the Entropic Graph Estimate

Let $\mathcal{S}^0 = \{\mathbf{s}_1^0, \ldots, \mathbf{s}_N^0\}$ be a set of N samples in $[0,1]^d$ and $\mathbf{u}_d$ be a unit vector in $\mathbb{R}^d$. Define $\mathcal{G}^*(\mathcal{S}^0) \triangleq \{G^*(\mathcal{S}^0)\}$, the set of all minimal graphs on $\mathcal{S}^0$. The following lemma states that after a slight perturbation of the value of a sample (within a certain limit) in $\mathcal{S}^0$, some of the current minimal graphs remain as minimal graphs and no other graph can become a minimal graph.

Lemma E.0.3. *For any $k \in \{0, \ldots, m\}$, there exists an $\epsilon > 0$ such that $\mathcal{G}^*(\{\mathbf{s}_1^0, \ldots, \mathbf{s}_k^0 + h\mathbf{u}_d, \ldots, \mathbf{s}_N^0\}) \subset \mathcal{G}^*(\mathcal{S}^0)$, for all $|h| \le \epsilon$.*

Proof. Let $\delta \triangleq \min_{G \in \mathcal{G}/\mathcal{G}^*}(W_\gamma(G(\mathcal{S}^0)) - W_\gamma^*(\mathcal{S}^0))$. Note $\gamma > 0$. If $|h| \le (\|e\|^\gamma + \delta/2N)^{1/\gamma} - \|e\|$ for all $\|e\|$ in G, then using the triangle inequality on each edge, it is easy to show that the change in $W_\gamma(G)$ is upper bounded by $\delta/2$. Recall that $\|e\| < \sqrt{d}$, since all $\mathbf{s} \in [0,1]^d$. Set $\epsilon = \max((\delta/2N)^{1/\gamma}, (d^{\gamma/2} + \delta/2N)^{1/\gamma} - \sqrt{d})$. Then for $|h| < \epsilon$ and all $G_1, G_2 \in \mathcal{G}(\mathcal{S}^0)$, the change in $W_\gamma(G_1) - W_\gamma(G_2)$ will be upper bounded by δ. Thus if $G \notin \mathcal{G}^*(\mathcal{S}^0)$, G will not achieve a $W_\gamma(G)$ smaller than $W_\gamma^*(\mathcal{S}^0)$. □

Now, let's look at the partial derivative of a power weighted edge length, $\|e_{ij}\| \triangleq$

$\|\mathbf{s}_i - \mathbf{s}_j\|$:

$$\frac{\partial(\|e_{ij}\|^\gamma)}{\partial s_{ic}} = \begin{cases} \gamma\|e_{ij}\|^{\gamma-2}(s_{ic} - s_{jc}) & \text{if } \mathbf{s}_i \neq \mathbf{s}_j, \\ 0 & \text{if } \mathbf{s}_i = \mathbf{s}_j \text{ and } \gamma \geq 1, \\ \pm\infty & \text{if } \mathbf{s}_i = \mathbf{s}_j \text{ and } \gamma < 1, \end{cases}$$

for $i, j = 1 \ldots N$ and $c = 1 \ldots d$. Note that, the derivative does not exist if the samples are coinciding and $\gamma < 1$. Elsewhere, it is well-defined.

The following lemma states the necessary and sufficient condition for W_γ^* to be differentiable.

Lemma E.0.4. *For a* $\mathbf{s}_k \in \mathcal{S}$*,* $\nabla_{\mathbf{s}_k} W_\gamma^*(\mathcal{S})$ *exists if and only if* $\nabla_{\mathbf{s}_k} W_\gamma(G^*(\mathcal{S}))$ *exists and is equal for all* $G^*(\mathcal{S})$*.*

Proof. Using the formal definition of the right derivative:

$$\begin{aligned} \partial W_\gamma^*(\mathcal{S})/\partial s_{kc}|_{s_{kc}=s_{kc}^{0+}} &= \lim_{h\to 0^+} \frac{W_\gamma^*(\{\mathbf{s}_1^0, \ldots, \mathbf{s}_k^0 + hu_{dc}, \ldots, \mathbf{s}_N^0\}) - W_\gamma^*(\mathcal{S}^0)}{h} \\ &= \min_{G\in\mathcal{G}^*(\mathcal{S}^0)} \partial W_\gamma(G)/\partial s_{kc}|_{s_{kc}=s_{kc}^{0+}}. \end{aligned} \tag{E.1}$$

Similarly the left derivative is equal to the maximum of the left derivatives among all the $G^*(\mathcal{S}^0)$'s. Now, consider the two cases:

1. $\mathbf{s}_k$ has a unique value $\mathbf{s}_k^0$. Then, $\nabla_{\mathbf{s}_k} W_\gamma(G^*(\mathcal{S}^0))$ exists for all $G^*(\mathcal{S}^0)$. Here, $\partial W_\gamma^*(\mathcal{S})/\partial s_{kc}$ exists if and only if the maximum and minimum derivatives are equal for all $c \in \{1, \ldots, d\}$.

2. $\mathbf{s}_k$ is not unique, i.e., there are other samples with the same value. Then it is easy to see that all minimal spanning graphs $G^*(\mathcal{S}^0)$'s contain at least one zero length edge with $\mathbf{s}_k$ as an endpoint. If $0 < \gamma < 1$, then the right and left derivatives of this edge length are $+\infty$ and $-\infty$, respectively. Thus $\nabla_{\mathbf{s}_k} W_\gamma^*(\mathcal{S})$

does not exist. If $\gamma > 1$, the edge length derivatives exist and the argument from 1 holds.

$\square$

Appendix F

Computing the EMST in 2D

In our implementation, to compute the EMST of a set of planar points V, we use Kruskal's algorithm preceded by a Delaunay triangulation.

Delaunay triangulation [20] of V, denoted by $DT(V)$, is the triangulation of V such that no vertex lies in the circumcircle of any of the triangles. It is known that the Delaunay triangulation is the geometric dual of the Voronoi tessellation. In 2D, a divide and conquer strategy yields a fast algorithm of $\mathcal{O}(N \log N)$ computational complexity, where N is the number of vertices [31].

Note that, like the MST, $DT(V)$ is not unique. However, an important *(circle) property* of the Delaunay triangulation is the following: If one can draw a circle with two vertices $v_1, v_2 \in V$ on its boundary, that contains no other vertices, then the edge $e(v_1, v_2)$ that connects v_1 and v_2 is in all $DT(V)$'s.

Kruskal's algorithm is a greedy, general purpose MST algorithm that computes the minimum spanning forrest (MSF) of an input graph [14]. If the graph is connected, the MSF is a MST. The pseudo-code for the algorithm is:

1. Create an empty set tree T that will hold the edges of the output forrest.

2. Create a *sorted* (by edge weight) edge set K that contains all edges of the input graph.

3. While K is not empty:

   (a) Remove the edge e with minimum weight from K,

   (b) If e creates no cycles in T, i.e., connects two disconnected trees, then add e to T.

The computational complexity of this algorithm is $\mathcal{O}(M \log N)$, where M and N are the number of edges and vertices in the input graph. If the input graph is complete, then the worst case complexity is $\mathcal{O}(N^2 \log N)$.

From Kruskal's algorithm, it is trivial to show the following *(cycle) property* of the MST: Within a set of edges that constitute a cycle, the edge with the most weight (if it exists) is not in any of the MST's.

The following, widely used result leads to an efficient 2D EMST algorithm.

Lemma F.0.5. *The edges in an EMST of a vertex set V is a subset of $DT(V)$.*

Proof. The proof trivially follows from the circle property of the Delaunay triangulation and cycle property of the EMST: Consider an edge e in the EMST of V, that connects v_1 and v_2. Assume e is not in any of the $DT(V)$'s. Draw the circle C, that has e as its diameter. Then, by the circle property, there exists at least one point v_3 in C. However, since e is the largest edge in the (v_1, v_2, v_3) cycle, by the cycle property, it cannot be in any of the EMST's of V. Thus, we have a contradiction. □

Hence, we can run Kruskal's algorithm on the Delaunay triangulation, which yields a computational complexity of $\mathcal{O}(N \log N)$ in 2D. In higher dimensions, the Delaunay triangulation (and thus the EMST algorithm) has worst case $\mathcal{O}(N^2)$ complexity.

Bibliography

[1] A. Antos and I. Kontoyiannis, "Convergence properties of functional estimates for discrete distributions," *Random Structures and Algorithms*, no. 19, pp. 163–193, 2001.

[2] B. Argall, Z. Saad, and M. Beauchamp, "Simplified intersubject averaging on the cortical surface using suma," *Human Brain Mapping*, vol. 27, no. 1, pp. 14–27, Jan. 2006.

[3] J. Beirlant, E. Dudewicz, L. Gyorfi, and E. van der Meulen, "Nonparametric entropy estimation: an overview," *International Journal Math. Stat. Sci.*, vol. 6, pp. 17–39, 1997.

[4] V. Blanz and T. Vetter, "Face recognition based on fitting a 3d morphable model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 1063–1074, 2003.

[5] F. Bookstein, "Principal warps: thin-plate splines and the decomposition of deformations," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 11, no. 6, 1989.

[6] L. Brown, "A survey of image registration techniques," *ACM Computing Surveys*, vol. 24, no. 4, pp. 325–372, 1992.

[7] T. Butz, O. Cuisenaire, and J. Thiran, "Multi-modal medical image registration: From information theory to optimization objective," *Proc. of DSP 2002*, pp. 407–414, 2002.

[8] T. Butz and J. Thiran, "Affine registration with feature space mutual information," *Proceedings of MICCAI 2001*, pp. 549–556, 2001.

[9] Y. Chaubey, G. Mudholkar, and P. Smethurst, "On entropy-based goodness-of-fit tests: a practical strategy," S. Basu and B. Sinha, Eds. New Delhi: Narosa Publishing House, 1993, pp. 116–120.

[10] G. Christensen, "Deformable shape models for anatomy," Ph.D. dissertation, Sever Institute of Technology, Washington University, 1994.

[11] A. Chung, W. Wells, A. Norbash, and W. Grimson, "Multi-modal image registration by minimising kullback-leibler distance," *Proc. of MICCAI'02*, 2002.

[12] C. Cocosco, V. Koolokian, R. Kwan, and A. Evans, "Brainweb: Online interface to a 3d mri simulated brain database," *NeuroImage*, vol. 5, no. 4, 1997.

[13] A. Collignon, D. Vandermeulen, P. Seutens, and G. Marchal, "3d mulit-modality medical image registration using feature space clustering," N. Ayache, Ed. Lecture Notes in C.Sci., Springer-Verlag, 1995, vol. 905, pp. 195–204.

[14] T. Corman, C. Leierson, R. Rivest, and C. Stein, *Introduction to Algorithms.* MIT Press and McGraw-Hill, 2001.

[15] J. Costa and A. Hero, "Geodesic entropic graphs for dimension and entropy estimation in manifold learning," *IEEE Trans. on Signal Processing*, vol. 52, no. 8, pp. 2210–2221, Aug. 2004.

[16] T. Cover and J. Thomas, *Elements of Information Theory.* Wiley Series in Telecommunications, 1991.

[17] D. Cremers, C. Guetter, and C. Xu, "Nonparametric priors on the space of joint intensity distributions for nonrigid multi-modal image registration," *Proc. of CVPR 2006*, 2006.

[18] E. D'Agostino, J. Modersitzki, F. Maes, A. Vandermeulen, B. Fischer, and P. Seutens, "Free-form registration using mutual information and curvature regularization," *Biomedical Image Registration*, vol. 2717, 2003.

[19] C. Davatzikos, J. Prince, and R. Bryan, "Image registration based on boundary mapping," *IEEE Trans. on Medical Imaging*, vol. 15, no. 1, Feb. 1996.

[20] B. Delaunay, "Sur la sphere vide," *Izvestia Akademii Nauk SSSR*, vol. 7, pp. 793–800, 1934.

[21] R. Duda, P. Hart, and D. Stork, *Pattern Classification.* Wiley-Interscience, 2001.

[22] E. Dudewicz, W. Mommaerts, and E. van der Meulen, "Maximum entropy methods in modern spectroscopy: a review and an empiric entropy approach," A. Öztrük and E. C. van der Meulen, Eds. Columbus, Ohio: American Sciences Press, Inc., 1991, pp. 115–160.

[23] E. Dudewicz and E. van der Meulen, "Entropy-based statistical inference, ii: Selection-of-the-best/complete ranking for continuous distributions on (0,1), with applications to random number generator," *Statistics and Decisions*, vol. 1, pp. 131–145, 1983.

[24] E.D'Agostino, F. Maes, D. Vandermeulen, and P. Suetens, "A viscous fluid model for multimodal non-rigid image registration using mutual information," *Medical Image Com. and Comp.-Ass. Int.*, vol. 2489, pp. 541–548, 2002.

[25] D. Erdogmus, "Information theoretic learning: Renyi's entropy and its applications to adaptive system training," Ph.D. dissertation, University of Florida, 2002.

[26] D. Erdogmus and J. Principe, "Information transfer through classifiers and its relation to probability of error," *Neural Networks, 2001. Proc. IJCNN '01*, vol. 1, pp. 50–54, July 2001.

[27] ——, “Lower and upper bounds for misclassification probability based on renyi’s information,” *Journal of VLSI Signal Proc. Systems*, vol. 37, no. 2/3, pp. 305–317, 2004.

[28] R. Fano, *Transmission of Information: A Statistical Theory of Communication.* New York, NY: MIT Press & John Wiley & Sons, Inc., 1972.

[29] B. Fischl, M. Sereno, and A. Dale, “Cortical surface-based analysis ii: Inflation, flattening, and a surface-based cooridnate system,” *Neuromimage*, vol. 9, pp. 195–207, 1999.

[30] B. Fischl, M. Sereno, R. Tootell, and A. Dale, “High-resolution inter-subject averaging and a surface-based coordinate system,” *Human Brain Mapping*, vol. 8, pp. 272–284, 1999.

[31] T. Q. A. for Convex Hulls, *ACM Transactions on Mathematical Software*, vol. 22, no. 4, pp. 469–483, Dec. 1996.

[32] L. Györfi and E. van der Meulen, “Denstiy-free convergence properties of various estimators of entropy,” *Computer Statist. Data Anal.*, vol. 5, pp. 425 – 436, 1987.

[33] J. Hajnal, D. Hill, and D. Hawkes, Eds., *Medical image registration.* CRC Press.

[34] A. Hamza and H. Krim, “Image registration and segmentation by maximizing the jensen-rényi divergence,” *EMMCVPR 2003, LNCS 2683*, pp. 247–263, 2003.

[35] U. Hasson, Y. Nir, I. Levy, G. Fuhrmann, and R. Malach, “Intersubject synchronization of cortical activity during natural vision,” *Science*, no. 303, pp. 1634–1640, 2004.

[36] J. Haxby, M. Gobbini, M. Furey, A. Ishai, J. Schouten, and P. Pietrini, “Distributed and overlapping representations of faces and objects in ventral temporal cortex,” *Science*, no. 293, pp. 2425–2429, 2001.

[37] M. Hellman and J. Raviv, “Probability of error, equivocation, and the chernoff bound,” *IEEE Transactions on Information Theory*, vol. 16, pp. 368–372, 1970.

[38] G. Hermosillo, C. Chefd'hotel, and O. Faugeras, "Variational methods for multimodal image matching," *Int. Journal of Computer Vision*, vol. 50, no. 3, pp. 329–343, 2002.

[39] A. Hero, J. Costa, and B. Ma, "Asymptotic relations between minimal graphs and α-entropy," *Technical Report CSPL-334 Communications and Signal Processing Laboratory*, Mar. 2003.

[40] A. Hero, B. Ma, O. Michel, and J. Gorman, "Applications of entropic spanning graphs," *IEEE Signal Proc. Mag.*, vol. 19, no. 5, pp. 85–95, 2002.

[41] A. Hero and O. Michel, "Asymptotic theory of greedy approximations to minimal k-point random graphs," *IEEE Transactions on Information Theory*, vol. 45, no. 6, pp. 1921–1938, 1999.

[42] C. Kuglin and D. Hines, "The phase correlation image alignment method," *Proc. IEEE 1975 Int. Conf. on Cybernetics and Society*, pp. 163–165, Sept. 1975.

[43] S. Kullback and R. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, no. 1, Mar. 1951.

[44] J. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, "Convergence properties of the nelder-mead simplex method in low dimensions," *SIAM Journal of Optimization*, vol. 9, no. 1, pp. 112–147, 1998.

[45] M. Leventon and W. Grimson, "Multi-modal volume registration using joint intensity distribution," *Proc. of MICCAI '98*, 1998.

[46] S. Li, *Markov Random Field Modeling in Computer Vision.* Springer-Verlag, 1995.

[47] B. Likar and F. Pernuš, "A hierarchical approach to elastic registration based on mutual information," *Image Vis. Computing*, vol. 19, no. 1-2, pp. 33–44.

[48] J. Lin, "Divergence measures based on shannon's entropy," *IEEE Trans. Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.

[49] B. Ma, A. Hero, J. Gorman, and O. Michel, “Image registration with minimum spanning tree algorithm,” *Proc. of ICIP ’00*, vol. 1, pp. 481–484, 2000.

[50] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Seutens, “Multimodality image registration by maximization of mutual information,” *IEEE Trans. on Medical Imaging*, vol. 16, no. 2, pp. 187–198, 1997.

[51] F. Maes, D. Vandermeulen, and P. Suetens, “Comparative evaluation of multiresolution optimization strategies for multimodality image registration by maximization of mutual information,” *Med. Image Anal.*, vol. 3, no. 4, pp. 373–386, 1999.

[52] J. Maintz and M. Viergever, “A survey of medical image registration,” *Medical Image Aanalysis*, vol. 2, no. 1, pp. 1–36.

[53] E. Miller, “Learning from one example in machine vision by sharing probability densities,” Ph.D. dissertation, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 2002.

[54] ——, “A new class of entropy estimators for mult-dimensional densities,” *Int. Conf. on Acoust., Speech and Signal Processing*, 2003.

[55] E. Miller, N. Matsakis, and P. Viola, “Learning from one example through shared densities on transforms,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 464–471, 2000.

[56] J. Modersitzki, “Numerical methods for image registration,” *Oxford University Press*, 2004.

[57] C. Moonen and P. Bandettini, *Functional MRI.* Springer-Verlag, 2000.

[58] V. Mountcastle, “The columnar organization of the neocortex,” *Brain*, vol. 120, pp. 701–722, 1997.

[59] H. Neemuchwala and A. Hero, “Entropic graphs for registration,” in *Multi-sensor Image Fusion and its Applications*, R. Blum and Z. Liu, Eds., 2004.

[60] H. Neemuchwala, A. Hero, and P. Carson, "Image matching using alpha-entropy measures and entropic graphs," *Signal Processing*, vol. 85, no. 2, 2002.

[61] M. Nielsen, P. Johansen, A. Jackson, and B. Lautrup, "Brownian warps: A least committed prior for nonrigid registration," *Medical Image Computing and Computer-Assisted Intervention*, vol. 2489, pp. 557–564, 2002.

[62] K. Nigam, J. Lafferty, and A. McCallum, "Using maximum entropy for text classification," *Proc. of the Int. Joint Conf. on Artif. Int., Workshop on Machine Learning for Information Filtering*, pp. 61–67, 1999.

[63] M. Ogawa, T. Miyata, K. Nakajima, K. Yagyu, M. Seike, K. Ikenaka, H. Yamamoto, and K. Mikoshiba, "The reeler gene-associated antigen on cajal-retzius neurons is a crucial molecule for laminar organization of cortical neurones," *Neuron*, vol. 14, pp. 899–912.

[64] M. Otte, "Elastic registration of fmri data using bezier-sline transformations," *IEEE Trans. Med. Imag.*, vol. 20, pp. 193–206, 2001.

[65] S. Ourselin, A. Roche, S. Prima, and N. Ayache, "Block matching: A general framework to improve robustness of rigid registration of medical images," *Medical Image Computing and Computer-Assisted Intervention*, vol. 1935, pp. 557–566, 2004.

[66] J. Pluim, J. Maintz, and M. Viergever, "Image registration by maximization of combined mutual information and gradient information," *IEEE Trans. Med. Imag.*, vol. 19, no. 8, pp. 809–814, 2000.

[67] ——, "Interpolation artefacts in mutual information based image registration," *Computer Vision and Image Understanding*, vol. 77, pp. 211–232, 2000.

[68] ——, "Mutual information based registration of medical images: A survey," *IEEE Trans. on Medical Imaging*, vol. 22, no. 8, pp. 986–1004, 2003.

[69] J. Principe, D. Xu, and J. Fisher, "Information theoretic learning," in *Unsupervised Adaptive Filtering*, S. Haykin, Ed. John Wiley & Sons, 2000.

[70] J. Rademacher, V. C. Jr, H. Steinmetz, and A. Galaburda, “Topographical variation of the human primary cortices: impilcations for neuroimaging, brain mapping and neurobiology,” *Cerebral Cortex*, vol. 3, pp. 313–329, 1995.

[71] P. Rakic, “Specification of cerebral cortical areas,” *Science*, vol. 241, no. 4862, pp. 170 – 176, 1998.

[72] C. Redmond and J. E. Yukich, “Asymptotics for euclidean functionals with power weighted edges,” *Stochastic Processes and their Applications*, vol. 6, pp. 289–304, 1996.

[73] A. Rényi, “On measures of information and entropy,” *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability 1960*, pp. 547–561, 1961.

[74] A. Roche, X. P. anc G. Malandain, and N. Ayache, “Rigid registration of 3-d ultrasound with mr images: A new approach combining intensity and gradient information,” *IEEE Trans. on Medical Imaging*, vol. 20, no. 10, pp. 1038–1050, Oct. 2001.

[75] A. Roche, G. Malandain, and N. Ayache, “Unifying maximum likelihood approaches in medical image registration,” *Int. Journal of Imaging Systems and Technology*, vol. 11, no. 1, pp. 71–80, May 2000.

[76] T. Rohlfing and C. Maurer, “Intensity-based non-rigid registration using adaptive multilevel free-form deformation with an incompressibility constraint,” *Medical Image Computing and Computer-Assisted Intervention*, vol. 2208, pp. 111–119, 2001.

[77] C. Rorden and M. Brett, “Stereotaxic display of brain lesions,” *Behavioural Neurology*, vol. 12, pp. 191–200.

[78] D. Rueckert, M. Clarkson, D. Hill, and D. Hawkes, “Nonrigid registration using higher order mutual information.” Bellingham, WA: SPIE Press, 2000.

[79] D. Rueckert, C. Hayes, C. Studholme, P. Summers, M. Leach, and D. J. Hawkes, “Nonrigid registration of breast mr images using mutual information,” in *MICCAI’98: Proceedings*. Springer-Verlag GmbH, 1998, vol. 1496.

[80] D. Rueckert, L. Sonoda, C.Hayes, D. Hill, and M. Leach, "Nonrigid registration using free-form deformations: application to breast mr images," *IEEE Transactions on Medical Imaging*, vol. 18, no. 1, pp. 712–722, 1999.

[81] M. Sabuncu and P. Ramadge, "Spatial information in entropy-based image registration," in *Biomedical Image Registration, LNCS 2717*. Springer-Verlag, 2003.

[82] ——, "Gradient based nonuniform sampling for information theoretic alignment methods," *Proc. of 26th Int. Conf. of the IEEE Engineering in Medicine and Biology*, Sept. 2004.

[83] ——, "Gradient based optimization of an emst registration function," *Proc. IEEE ICASSP '05*, Mar. 2005.

[84] ——, "Graph theoretic image registration using prior examples," *Proc. of EUSIPCO '05*, Sept. 2005.

[85] C. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423,623–656, 1948.

[86] S. Smith, "Fast robust automated brain extraction," *Human Brain Mapping*, vol. 17, no. 3, pp. 143–155, Nov. 2002.

[87] R. Stoica, J. Zerubia, and J. Francos, "Image retrieval and indexing: A hierarchical approach in computing the distance between textured images," *IEEE Int. Conf. on Image Processing*, 1998.

[88] C. Studholme, D.L.G.Hill, and D. Hawkes, "An overlap invariant entropy measure of 3d medical image alignment," *Pattern Recognition*, no. 1, pp. 71–86.

[89] C. Studholme, D. Hill, and D. Hawkes, "Multiresolution voxel similarity measures for mr-pet registration," Y. Bizais, C. Barillot, and R. Paola, Eds.

[90] J. Talairarch and P. Tournoux, *Co-planar Stereotaxic Atlas of the Human Brain: 3-Dimensional Proportional System - an Approach to Cerebral Imaging.* New York, NY: Thieme Medical Publishers, 1988.

[91] P. Thvenaz and M. Unser, “Optimization of mutual information for multiresolution image registration,” *IEEE Trans. on Image Proc.*, vol. 9, no. 12, pp. 1083–1100, 2000.

[92] R. Tootell, J. Reppas, K. Kwong, R. Malach, R. Born, T. Brady, B. Rosen, and J. Belliveau, “Functional analysis of human mt and related visual cortical areas using magnetic resonance imaging,” *J Neuroscience*, vol. 15, pp. 3215–3230, 1995.

[93] P. van den Elsen, E. Pol, and M. Viergever, “Medical image matching a review with classification,” *Engineering in medicine and biology*, vol. 12.

[94] D. Vandermeulen, A. Collignon, J. Michiels, H. Bosmans, P. Suetens, G. Marchal, G. Timmens, P. van den Elsen, M. Viergever, H. Ehricke, D. Hentschel, and R. Graumann, “Multi-modality image registration within covira,” *Studies in health,technology and informatics*, vol. 19, pp. 29–42, 1995.

[95] S. Verdú, “Fifty years of shannon theory,” *IEEE Trans. on Info. Theory*, vol. 44, no. 6, pp. 2057–2079, Oct. 1998.

[96] P. Viola, “Alignment by maximization of mutual information,” Ph.D. dissertation, Dept. of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1995.

[97] P. Viola and W. Wells, “Alignment by maximization of mutual information,” *Int. Journal of Computer Vision*, vol. 24, no. 2, pp. 137–154, 1997.

[98] R. Woods, S. Cherry, and J. Maziotta, “Rapid automated algorithm for aligning and reslicing pet images,” *Journal of Comput Assist Tomogr.*, vol. 16, no. 4, pp. 622–633, 1992.

[99] K. Worsley and K. Friston, “Analysis of fmri time-series revisited - again,” *NeuroImage*, no. 2, pp. 173–181, 1995.

[100] Y.He, A. Hamza, and A. Krim, “A generalized divergence measure for robust image registration,” *IEEE Trans. Signal Processing*, vol. 51, no. 5, pp. 1211–1220, May 2003.

[101] J. Yukich, *Probability Theory of Classical Euclidean Optimization.* Berlin: Springer-Verlag, Lecture Notes in Computer Science, 1998, vol. 1675.

[102] L. Zöllei, J. Fisher, and W. Wells, "An introduction to statistical methods of medical image registration," in *Handbook of Mathematical Models in Computer Vision*, N. Paragios, Y. Chen, and O. Faugeras, Eds. Springer, 2006, pp. 531–542.

[103] L. Zollei, J. Fisher, and W. Wells, "A unified statsitical and information theoretic framework for multi-modal image registration," *IPMI 2003, LNCS 2732*, pp. 366–377, 2003.

[104] L. Zöllei, E. Learned-Miller, E. Grimson, and W. Wells, "Efficient population registration of 3d data," *Proceedings of ICCV*, 2005.