Unifying Models and Registration: A Framework for Model-based Registration and Non-rigid Registration Assessment

A thesis submitted to the University of Manchester

for the degree of Doctor of Philosophy in the

Faculty of Medical and Human Sciences

2007

Roy Samuel Schestowitz

Division of Imaging Science and Biomedical Engineering

Contents

1 Introduction **22**

1.1 Image Registration 23

1.2 Statistical Appearance Models 25

1.3 Exploiting the Connection 26

1.4 Contributions 27

1.5 Thesis Organisation 28

2 Non-rigid Registration **30**

2.1 Motivation 31

2.2 NRR Methods 32

2.2.1 Overview 32

2.2.2 Representing Image Deformation 33

2.2.3 Objective Function 39

2.2.4 Optimisation 44

2.2.5 Pairwise versus Groupwise Registration 45

2.3 Assessment of NRR 47

2.3.1 Recovery of Deformation Fields 47

2.3.2 Overlap-Based Assessment 48

2.3.3 Motivation for Ground-Truth-Free Assessment 50

2.4 Summary 51

3 Active Appearance Models **53**

3.1 Motivation 54

3.2 Shape Models 56

3.2.1 Representing Shapes 57

3.2.2 Aligning Shapes 58

3.2.3 Obtaining Modes of Variance 59

3.2.4 The Correspondence Problem 62

3.3 Appearance Models 63

3.3.1 Modelling Shape 64

3.3.2 Modelling Intensity 64

3.3.3 Combined Models 66

3.4 Active Appearance Models 69

3.4.1 Learning to Fit the Model 70

3.4.2 Model Fitting 72

3.5 Summary 75

4 MDL Shape Models **76**

4.1 Overview 77

4.2 Measuring Model Quality 79

4.3 Methods of Manipulating Correspondence 80

4.4 Optimisation 83

4.5 Extensions 85

4.6 Summary 90

5 Model-based Registration **91**

5.1 Overview 92

5.2 Applying the MDL Approach to Images 93

5.2.1 Image Data 93

5.2.2 Experimental Framework 97

5.2.3 Image Warping 98

5.2.4 Objective Function 100

5.2.5 Reparameterisation and Optimisation 101

5.2.6 The Registration Algorithm 102

5.2.7 The Algorithm Visualised 104

5.3 Initial Experiments - Registering Images 105

5.3.1 Model-based Registration Experiments 105

5.3.2 Comparison of Registration Methods 108

5.3.3 Deficiencies of MDL-based Models 111

5.3.4 Comparison of Optimisation Regimes 112

5.4 Experiments 112

5.4.1 Weighting 113

5.4.2 Optimiser Tolerance Section 116

5.4.3 Subsets 119

5.5 From Registration to Models 124

5.5.1 Automatically Building Appearance Models in 1-D 125

5.5.2 2-D Model Construction 128

5.6 Summary and Discussion 133

6 Assessment of Non-Rigid Registration **135**

6.1 Overview . . . 136

6.2 Measuring Model Quality . . . 137

6.2.1 Specificity . . . 137

6.2.2 Generalisation . . . 140

6.2.3 Caveat . . . 142

6.3 Image distance measures . . . 143

6.4 Summary and Discussion . . . 145

7 Validation Methodology and Experiments **148**

7.1 Method . . . 149

7.1.1 Image Data . . . 149

7.1.2 Perturbing the Initial Registration . . . 150

7.1.3 Validation using Warped Images . . . 152

7.1.4 Sensitivity . . . 154

7.2 Results . . . 157

7.2.1 Sensitivity . . . 158

7.2.2 Effect of Noise . . . 159

7.3 Discussion . . . 159

8 Application to Evaluation of NRR **162**

8.1 Comparing Registration Algorithms . . . 163

8.1.1 Pairwise Registration to a Reference . . . 164

8.1.2 Groupwise Congealing Algorithm . . . 165

8.1.3 Groupwise MDL Algorithm . . . 166

8.2 Results of Comparison . . . 166

8.3 Exploration of 3-D NRR . . . 169

9 Future Work **170**

9.1 Automatic Model Building 170

9.2 Assessment of NRR 171

9.2.1 Specificity and Generalisation 171

9.2.2 Investigating Robustness 173

9.2.3 Extension to 3-D 173

10 Summary and Conclusions **177**

10.1 Discussion 178

10.2 Conclusions 181

Abstract

Statistical models of shape and appearance are widely used for analysis of biomedical images. Two deficiencies of these models are that they require consistent annotation of a large number of images in order to be built, and having built such models, it is then difficult to reason about their validity, let alone assess their quality. Herein, a method is described which addresses both problems and establishes a unified solution. In order to construct models reliably and rapidly, corresponding structures must be brought into a state where dense overlap across images is obtained. Image registration is the mechanism whereby a set of images can be analysed in a common frame of reference and models then derived from it. The thesis provides a solution to the problem where there is a recurring need to compare such models. It extends the method as to provide an image registration assessment method which does not require ground-truth data. The thesis also deals with a complementary case where images are registered by minimising the complexity of models. Overall, the proposed framework can be perceived as one which combines registration and modelling, taking advantage of the fact that the ideas behind them

are inherently the same. Registration provides correspondence across images and, given that correspondence, models of appearance can be built and registration then assessed, without the need for ground-truth data.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

sion of Imaging Science and Biomedical Engineering.

Dedication

I dedicate this thesis to my grandmother, who passed away amidst my studies.

Acknowledgements

I would like to express my thanks to:

- My Supervisor and role model, Chris Taylor, for investing plenty of time and effort to make my project a success.
- My Advisor, Steve Williams, for support, kind words of advice and perpetual encouragement.
- Katherine Smith, Carole Twining, and Stephen Marsland, who initiated much of the work which the thesis relates to, as well as builds upon.
- Tim Cootes, who helped me cope with C++ and VXL (among other technicalities). He has always done so while remaining patient and keeping pressure (or a sense of demand) at a bare minimum.
- The EPSRC, for funding the project and connecting me with those who are involved in the MIAS-IRC. This 'umbrella' offered valuable feedback, advice, and a very motivational setting.

- My parents and siblings, as well as my grandfather, Avner Werner Max, for full financial support from my very first day at the University.

- My grandmother, who is no longer among us (passed away in 2005). She was a supportive and loving figure in a family which has been volatile for the past 2 years.

- Dr. David Baxter, whose advice led me to consider research at Manchester University and, more particularly, the division where I ended up enrolling for this degree.

- Prof. Tony Hegarty, who urged me to accept an academic route, rather than be distracted by the wrong career paths.

- David Robinson, who offered valuable advice and encouraging words towards the end of my Ph.D. He was one of those who urged me to put in extra time and rigour, gearing up towards completion.

- Other friends at the health club, who inspired and provided comfort whenever things went amiss or uncertainties loomed over.

- The Open Source community, which enabled this research to be carried out using versatile, distributable, Free software.

- Mathworks, for enabling me to share code in their Web site and thereby reach a very wide audience, putting to use some personal programming ambitions. Many of my endeavours were boosted by

knowledge that the outcome would be shared among the MATLAB users community.

- Roland Selby, Mark O'Leary and several other people at Manchester Computing, who permitted me to concentrate on my Ph.D. studies while at work. They were surprisingly understanding, particularly when pressure was mounting.

- David Kennedy of the Center for Morphometric Analysis at MGH, for providing the fully-annotated brain images. Additional images from age-matched normals in a dementia study were generously provided by Prof. Alan Jackson, University of Manchester.

Publications

Portions of the work described in this thesis have also appeared in:

Conference papers

- Carole Twining, Tim Cootes, Stephen Marsland, Vladimir Petrovic, Roy Schestowitz, and Chris Taylor. A Unified Information-Theoretic Approach to Groupwise Non-Rigid Registration and Model Building. Presented in *Information Processing in Medical Images (IPMI)*, Lecture Notes in Computer Science, vol. 3565, pp. 1–14, 2005.

- Tim Cootes, Carole Twining, Vladimir Petrovic, Roy Schestowitz, and Chris Taylor. Groupwise Construction of Appearance Models using Piece-wise Affine Deformations. Presented in *British Machine Vision Conference (BMVC)*, vol. 2, pp. 879–888, 2005.

- Roy schestowitz, Carole Twining, Tim Cootes, Vladimir Petrovic, Chris Taylor, and Bill Crum. Assessing the Accuracy of Non-Rigid

Registration With and Without Ground Truth. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2006.

- Carole Twining, Tim Cootes, Stephen Marsland, Vladimir Petrovic, Roy schestowitz, and Chris Taylor. Information-Theoretic Unification of Groupwise Non-Rigid Registration and Model Building. In *Medical Image Understanding and Analysis (MIUA),* vol. 2, pp. 226–230, 2006.

- Roy Schestowitz, Carole Twining, Tim Cootes, Vladimir Petrovic, Bill Crum, and Chris Taylor. Non-Rigid Registration Assessment Without Ground Truth. Presented in *Medical Image Understanding and Analysis (MIUA)*, vol. 2, pp. 151–155, 2006.

Peer-reviewed papers/symposia

- Roy Schestowitz, Carole Twining, Tim Cootes, and Chris Taylor. Image Registration by Model Criteria. Presented in *Proceedings of MIAS-IRC Plenary Meeting*, pp. 16–17, 2004.

- Roy Schestowitz, Bill Crum, Vladimir Petrovic, Carole Twining, Tim Cootes, and Chris Taylor. Assessing the Accuracy of Non-Rigid Registration. Presented in *Proceedings of MIAS-IRC Plenary Meeting*, pp. 25–26, 2005.

- Vladimir Petrovic, Tim Cootes, Carole Twining, Roy Schestowitz, and Chris Taylor. Groupwise Construction of Appearance Models

using Piece-wise Affine Deformations. Presented in *Proceedings of MIAS-IRC Plenary Meeting*, pp. 21–22, 2005.

- Roy Schestowitz, Carole Twining, Tim Cootes, Vladimir Petrovic, and Chris Taylor. A Generic Method for Evaluating Appearance Models. Presented in *Proceedings of MIAS-IRC Plenary Meeting*, pp. 48–49, 2006.

- Vladimir Petrovic, Tim Cootes, Carole Twining, Roy Schestowitz, and Chris Taylor. Simultaneous Image Registration, Segmentation, and Modelling. Presented in *Proceedings of MIAS-IRC Plenary Meeting*, pp. 46–47, 2006.

Journals (submitted, under review)

- Roy Schestowitz, Carole Twining, Vladimir Petrovic, Tim Cootes, Bill Crum, and Chris Taylor. Evaluating Non-Rigid Registration without Ground Truth. *IEEE Transactions on Medical Imaging*.

Code and material produced throughout the project was generalised and publicised in the form of utilities and tutorials. It received over 60,000 downloads at MATLAB Central and had the author of this thesis ranked 1^{st} in the world at one point. Such work is now internationally recognised and it receives approximately 2,000 downloads per month.

Accompanying CD-ROM

Videos demonstrating parts of the thesis have been put on a supplementary CD-ROM. This CD-ROM is bound to the back cover of this thesis. It contains illustrative animations that are sometimes referenced in the text. The filenames on the CD-ROM are arbitrary. These are neither chronologically sorted, nor do they correspond to figure/section numbers in the text.

Peripheral files (GIF-formatted):

- /Thesis/p1.gif: a video of a combined model built from the MGH data. Seven modes of variation are normally (i.e. drawn from Gaussian distribution) varied simultaneously.
- /Thesis/p2.gif: combined model built from the MGH data. Ten modes of variation are normally varied simultaneously.
- /Thesis/p3.gif: the registration assessment framework illustrated schematically

- /Thesis/p4.gif: chequerboard image of 2 brain images. The sequence shows the chequerboard image as an SSD-based registration proceeds and, in response, the chequerboard image evolves.

- /Thesis/p5.gif: one-dimensional registration example. Rows in the matrix represent intensity vectors (1-D images) being registered to align with the remainder of the vector set in a pair-wise fashion.

Larger peripheral files:

(Compressed or uncompressed AVI format; no proprietary codecs required)

- /Thesis/p6.avi: a set of 10 1-D vectors are being aligned using multi-edge clamped plate spline (CPS) warps. Ten iterations (systematic passes) through the data are shown, unregistered images on the left and progressively re-registered images on the right.

- /Thesis/p7.avi: a set 10 1-D vectors, as visualised in 3-D space, are being aligned. Fifty iterations through the data are shown in the sequence.

- /Thesis/p8.avi: the first and second modes of a combined (shape and intensity) model which is automatically built

- /Thesis/p9.avi: two 1-D images are being registered. Unregistered images are shown on the left and progressively registered – on the right.

- /Thesis/p10.avi: 1-D vectors being registered using a model-based objective function
- /Thesis/p11.avi: 10 simplified 1-D vectors, which are composed of 4 edges, are being registered. The objective function is based on mean-squared-differences.
- /Thesis/p12.avi: 10 1-D vectors with are being registered using an objective function that based on minimisation of the complexity of a point distribution function
- /Thesis/p13.avi: A large number of 1-D vectors (visualised as rows) being registered by considering just one vector at a time
- /Thesis/p14.avi: automatically-built combined model of a bump. The model is built automatically from the raw training set.
- /Thesis/p15.avi: A large-scale illustration of 1-D bump registration

Prologue

"The classical *synthesis problem* of computer graphics can be formulated as the problem of generating novel images corresponding to an appropriate set of parameters describing the camera viewpoint and aspects of the scene. The inverse *analysis problem* of estimating object labels as well as scene parameters from images is the classical problem of computer vision..."

– *David Beymer* [8].

Chapter 1

Introduction

"A mathematician is a device for turning coffee into theorems."

– *Paul Erdos.*

THIS thesis studies the relationship between models and registration. The main contribution is a novel approach to the evaluation of statistical models of appearance [31], which can also be used to assess the quality of non-rigid registration (NRR) algorithms. Additionally, a method is presented for registering images, using model complexity as a figure of merit [94].

The work is motivated by the observation that, given a set of registered (i.e. fully-aligned) images [41], a statistical model of appearance can be built automatically. This is exploited in two ways: first, we propose an objective function for automatic NRR of sets of images, based on an

information-theoretic measure of model complexity; second, we show how the quality of NRR of a set of images can be assessed by measuring the quality of the resultant model. Both are important, but the thesis makes a more significant contribution to the latter. The ability to assess NRR algorithms is important for benchmarking and comparative studies, or for quality control in practical applications.

This work thus makes a contribution to both the modelling and non-rigid image registration literature. The approach is generic, but the thesis is focused on 2-D brain images. The extension to 3-D is discussed and, at the 'proof-of-concept' stage, 1-D images are also used. These are helpful in validation experiments that exploit synthetic images whose nature is well understood.

1.1 Image Registration

As medical imaging methods have developed rapidly over the past decade, the ability to discover clinically significant patterns in large quantities of data has become important. One class of problems involves comparing groups of images. There are two common cases:

1. **Intra-subject studies** involve analyzing a series of images taken from the same subject and comparing them. These images may be acquired over a period of time, as means of tracking the progression

of abnormalities, e.g. [109]. Alternatively, they may be taken at roughly the same time with different image modalities, elucidating different properties of the tissue being imaged.

2. **Inter-subject studies** involve a comparison between two (or more) groups of subjects. In a medical context, this makes possible the discovery of symptoms for a certain groups of patients, observing how they deviate from 'normals'.

Fundamental to this is the need to compare images. Non-rigid registration (NRR) is commonly used to address problems of this kind [17, 28, 41]. NRR algorithms warp two or more images into the same frame of reference so that equivalent structures align. The problem is highly underconstrained and many different algorithms have been proposed for solving it with no consensus as to which particular algorithm should be used.

The aim of NRR is to find, automatically, a meaningful dense correspondence between a pair (*pairwise* registration), or across a group of images (*groupwise* registration). A typical algorithm requires a representation of the deformation fields that encode the spatial variation between images, an objective function that quantifies the degree of similarity between the registered images, and a method of optimising the objective function with respect to the deformation fields.

Since different algorithms generally produce different results when applied to the same set of images [134], there is a clear need for methods

that evaluate the results of NRR. This way, the differences between methods can be investigated and – for example – quality control (QC) can be exercised on a case-by-case basis in clinical studies.

Various methods of evaluation have been proposed [46, 87, 96] for assessment of the results of NRR. One approach is to construct artificial test data and then apply known deformations to real or synthetic images. This allows algorithms to be evaluated by attempting to recover the applied deformations, but this does not allow the results of NRR to be assessed for QC purposes in real applications. An alternative approach is to provide anatomical ground truth for the images to be registered, then measure the degree of anatomical correspondence following NRR. One such method is used in this thesis as a gold standard, but the need for expert annotation of the images renders the approach too time-consuming and subjective for routine application. These problems motivate the search for a method of evaluation that can be used routinely in real applications, without requiring ground truth. This can be achieved by exploring the relationship between NRR and statistical model building.

1.2 Statistical Appearance Models

Statistical models of shape and appearance have been used extensively as a basis for image interpretation by synthesis [20, 34]. The basic idea is

that, given a dense correspondence between the set of images of the same class of structures, it is possible to build coupled statistical models of the shape variation and shape-free texture variation across the set[1]. Given the generative 'appearance model', it is possible to interpret new images of the same type by finding an instance of the model that matches the image as closely as possible. There are efficient algorithms that achieve this by finding a best match [19, 23, 24].

1.3 Exploiting the Connection

There are two ways in which the connection between NRR and models can be exploited. NRR can be assisted by measures of model complexity and models can be built from NRR.

The main contribution of this thesis is the introduction of a generic method for assessing the quality of non-rigid registration [93]. The method *does not* require ground truth, but rather depends solely on the registered images. Consider the case where NRR is applied to a *set* of images, providing a dense correspondence between these images. Given this correspondence, it is possible to build a generative statistical model of appearance variation for the set. The quality of the resultant model will depend on the quality of the correspondence. The key idea that underpins the approach is that, if the correspondence is poor, the resulting appearance

[1] In this thesis, "texture" is used in the graphics sense to mean the intensity pattern.

model will be unsatisfactory. When the correspondences are correct, the model will be better. We define measures of model *specificity* and *generalisation,* which can be used to assess the quality of the model and, hence, the quality of the correspondence from which it is derived. This approach transforms the problem of assessing non-rigid registration to one of evaluating models generated from the results of registration. It does not depend on the specifics of the registration algorithm or the form of the model.

1.4 Contributions

The contribution of the work is two-fold:

- On the one hand, it is demonstrated that one is able to evaluate the quality of NRR without requiring ground-truth data. Thus, a comparison between various NRR algorithms can be made without cumbersome-to-obtain manual annotation.
    - ✧ Further to this, experimental validation is used to show that the method works.
    - ✧ Several registration algorithms are compared using the new method, demonstrating the use of this method in a practical setting.

- On the other hand, one is also able to build models automatically, using registration algorithms, and then evaluate the resultant models.

Overall, this provides a framework for automatic or semi-automatic analysis of arbitrarily-large amounts of data.

1.5 Thesis Organisation

The structure of the thesis is as follows:

Chapters 2 and 3 provide detailed descriptions of existing methods of NRR and constructing appearance models, respectively. They also explore the link between the two.

Chapter 4 discusses previous work on constructing shape models using a minimum description length (MDL) approach.

Chapter 5 builds on Chapter 4. It briefly outlines an approach, which combines NRR and appearance modelling, by using a model-based MDL objective function that drives registration. Results are shown for the authors' initial experiments in 1-D images, and from a collaborative project where 2-D images are used.

Chapter 6 exploits the duality of models and registration quality defining two quantitative measures of model/registration quality which can be used to assess the results of NRR without the need for ground truth.

Chapter 7 describes a series of validation experiments which investigate the effects of deliberately perturbing the registration of an initially registered set of images. The results obtained using the measures introduced in Chapter 6 are compared to those obtained using a 'gold standard' method of assessment, based on measuring the overlap of manually-annotated ground truth label images.

Chapter 8 presents a practical application of the new method of evaluating NRR. Three different NRR algorithms applied to the registration of sets of 2-D MR brain images, demonstrating the superiority of a fully groupwise registration algorithm over a repeated pairwise approach.

Chapter 9 lists several possible extensions and several ways forward. It also describes the existing extension of the method to 3-D, and discusses its limitations.

Chapter 10 draws conclusions and provides a summary of the contributions of the work.

Chapter 2

Non-rigid Registration

"Having a set, popular formula does inhibit you."

– *George Shearing.*

THIS chapter provides a detailed introduction to image registration. It explores methods that are used in NRR algorithms while discussing both their strengths and weaknesses. It explains the motivation for registration and then lists a few practical applications which benefit from it. The constituent parts of an NRR algorithm are detailed and dealt with in turn. Then, a comparison is drawn which rationalises the choice of one registration algorithm over another. The last part of this chapter deals with methods that can be used to assess the quality of NRR algorithms.

2.1 Motivation

First and foremost, what motivates registration is the need to compare images of objects. If the images (and hence the objects within them) can be properly aligned, then it is possible to study the differences between the imaged objects.

The importance of registration is demonstrated by its ubiquity. It has entered several domains where reliable acquisition of fully aligned images cannot be assured [41, 82] or relationships between images turn out to be complex. The difficulty of this problem is even greater when alignment of large *groups* of images needs to be achieved [119]. There is room for ambiguity – and consequently – misinterpretation. NRR brings together large amounts of data, enabling the examination of variation.

Misalignment in images can result from movement of subjects or objects of interest, change in view-point, or changes to general conditions at the acquisition. Misalignments can also be artifacts of morphological changes or physical abnormalities that are due to change in mass and elasticity of organs [42].

2.2 NRR Methods

2.2.1 Overview

The field which is associated with this problem is uniformly referred to in the literature as "non-rigid registration". The subject is broad and for deeper understanding of alternative approaches, there are useful text-books [41].

Given a collection of images, all of which depict the same type of object, one wishes to transform the images, so that the correspondingly transformed objects within the images now appear as similar as possible to one another. The solution to this problem is never a unique one as there will be infinitely many transformations, which lead to similar results. Since such images may not contain precisely the same elements, there is rarely an absolute one-to-one correspondence between imaged objects. Absence or reappearance of finer elements, for example, mean that no unique point-to-point correspondence can exist, so good solutions need be *approximated* instead.

The aim of non-rigid registration of medical images is to properly identify an anatomically-meaningful, dense (i.e., pixel-to-pixel or voxel-to-voxel) correspondence across a set of images. This correspondence is typically encoded as a set of spatial deformation fields, one for each image, such that when the deformations are applied to the images, corresponding structures are brought into alignment.

Typical registration algorithms proceed by optimising an objective function, which depends on the similarity of the images after alignment, with respect to the set of deformations [80]. As well as the objective function, it is necessary to define the representation used for the deformation fields and the method for finding the optimum of the objective function. The three components of an NRR algorithm are:

- Representation of image deformation
- Objective function
- Optimisation algorithm

Different choices of any of these elements lead to different registration results and competing methods of NRR – hence the need for an objective method of assessment. The following sections deal with each component in turn.

2.2.2 Representing Image Deformation

There are various image deformation methods which produce a full transformation that affects an entire image plane or volume [3, 69, 76]. Sometimes, the transformations (also known as "warps") are also subjected to conditions that ensure they remain valid. A transformation might, for example, be carried out under the imposition of strict constraints. These

prevent the transformation from collapsing upon itself – something which leads to problems that are explained later.

Transformations can be subdivided into different types. Each type of transformation has a different effect on the data it is applied to and a typical classification of transformation types is as follows (ordered by increasing level of versatility):

Euclidean Transformation. Permits translation (relocation in space), rotation, and scaling[1]. In space, normalised shape attributes are preserved, so it generally provides an approximate alignment. Such alignment is ordinarily intended to position all data instances (images or volumes) upright and centred at the origin of a space, with a fixed size of 1 unit at most. All images are confined to lie inside a bounding structure (for example, a circle in 2-D, or a sphere in 3-D). In 3-D, for instance, there is a total of 7 degrees of freedom so a rigid transformation [63] will be wholly characterised by a tuple of 7 parameters (scale included).

Affine Transformation. Allows an image to *stretch and skew* along at least one axis (corresponding to a single dimension). Despite the fact that dimensions can be rescaled at different magnitudes, lines which were parallel before an affine transformation is applied will

[1]The exclusion of scaling makes this a rigid transformation, rather than Euclidean transformation.

remain parallel after the transformation is applied[2]. This transformation is also invertible. Essentially, for a given affine transformation $T_a(x)$, where x is a vectorised representation of an image (or volume), and its inverse $T_a^{-1}(x)$, the expression $T(T_a^{-1}(x)) = Id(x)$, where $Id(x)$ is the identity operator, must hold as valid. It retains a level of simplicity, which makes it easy to determine and resolve. This proves to be an important constraint when the practicability of warps is further debated.

Non-rigid Transformation. There is a large variety of ways of transforming images more flexibly. They do not comply with the constraints imposed by Euclidean and affine transformations. Among the methods for transforming biomedical images, there is a fluid flow framework, which models deformations in terms of the flow of a fluid throughout the image region [15]. Hence, it potentially allows large-scale deformations. Clamped-plate splines [114], on the other hand, make use of the composition of rounded [64] and localised warps. When these warps are aggregated, they facilitate flexible deformation at multiple scales. Thin-plate splines [10] are another way of formulating the transformation of pixels. Another common approach, for instance, models the motion of organs, which can be handled using free-form deformations (FFDs). These can be based on B-splines [91].

[2]Other transformations such as taper, on the contrary, are not parallelism preserving. The importance of this rigorous constraint is that the distance between any two points remains proportional to the transformation.

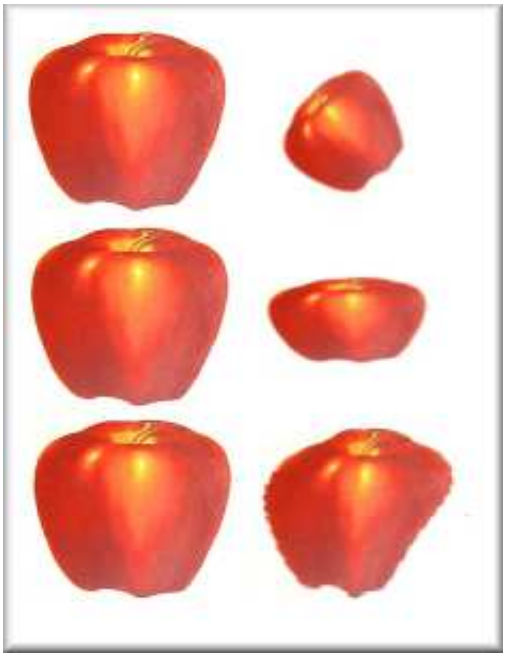

Figure 2.1: Registration examples. On the left column: the original, unwarped image; on the right column, from top to bottom: rigid, affine, and non-rigid transformations.

The images of an apple in Figure 2.1 illustrate the effect that each transformation type is permitted to have on the original image shown on the left.

As the figure suggests, the appearance of an object remains identical under rigid transformations. Imaged objects are allowed strictly to grow, shrink, move, and rotate. Affine transformation allows an object to lose its original form whereas non-rigid registration is far more permissive, so the object can be subjected to rather arbitrary deformations.

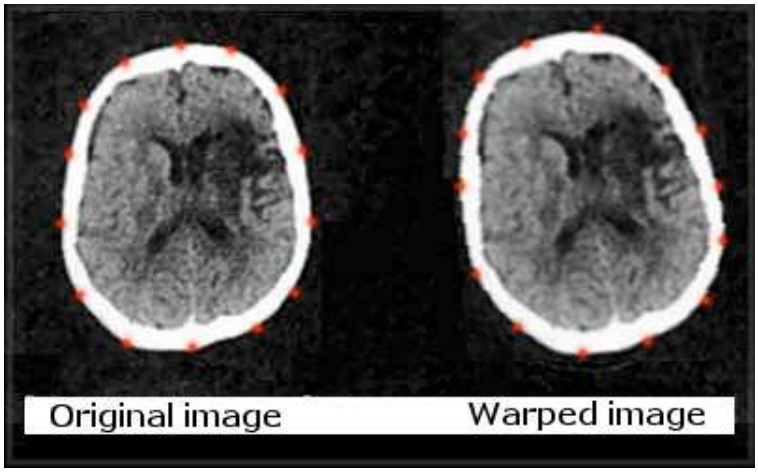

Figure 2.2: An example of image warping in medical contexts (the human brain). Points that are overlaid on the skull depict knot-points for the splines that render a transformation, which is based on clamped-plate splines. The image on the right is a warped and tilted version of the original on the left.

There are important factors to consider when selecting a transformation method. One such factor is the property called diffeomorphism. Diffeomorphic [112] functions are *invertible, continuous* and *one-to-one* mappings, which can be applied to a given image. Diffeomorphic transformations that are used in this work were initially devised by Twining and Marsland [114] (see the example in Figure 2.2). These benefit from having continuous derivatives at the boundaries unlike, for example, those proposed by Lötjönen and Mäkelä [64].

What invertibility, continuity and one-to-one mappings mean, in simpler terms, is that for each transformation:

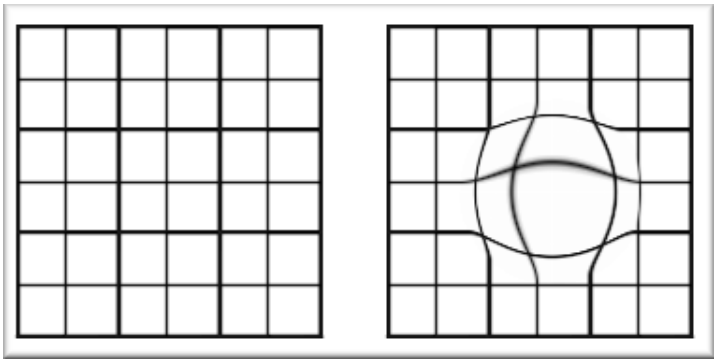

Figure 2.3: A pseudo-non-rigid warp example. The effect of the warp on a normal grid (left) is shown on the right hand side. Because lines in the grid do not intersect by collapsing onto one another, the warp is considered to be diffeomorphic.

1. The transformation has a calculable inverse transformation. This way, any transformation can in principle be reversed, i.e. its effect retracted.

2. The transformation affects *all* data (e.g. image pixels) within its boundaries so it has a spatially-contained effect[3]. This means that every point must move as would be expected to give a continuous flow of intensities.

3. No two points should be mapped onto the same point as this would 'strip off' areas of the image, depleting them from data.

These latter two effects are also known as tearing and folding, both of which are notorious artifacts that need to be avoided (see Figure 2.3).

[3]A pixel of course can be mapped onto the exact same original position, but the idea is that a continuous flow should remain.

2.2.3 Objective Function

The purpose of the objective function is to quantify the goal of NRR algorithms by introducing measurable attributes. This typically involves quantification of image dissimilarity, either by a direct pixel-by-pixel or voxel-by-voxel comparison, or by using less direct image histogram methods. Most of the methods below are area based, which means that the NRR algorithm analyses structures in the images via correlation metrics and other means of performing a structural analysis, e.g. Fourier properties. Feature-based registration methods are discussed as well, although the results presented later in this thesis do not make use of such methods and need not consider Euclidean distance measures *within* images, only between separate images in hyperspace.

I describe several variants of these two sorts of approaches in what follows. In essence, all are methods for establishing a measure of *distance* between images, but only the means vary.

Image Difference Measures

There are various ways of measuring a perceived similarity between two images and selection of a method depends on the images at hand [105] as well as dimensionality. Sum of pixel- or voxel-wise differences (also sum-of-*squared*-differences or even the *mean* rather than a summation) emerges as the most intuitive method, which merely accumulates pixel-

or voxel-wise differences between the images. This measure, however, performs rather poorly in situations where slight miscorrespondence results in very high localised differences. If areas of the images to be aligned are not smooth, sum-of-squared-differences will be insensitive to miscorrespondence. Moreover, sum-of-squared-differences is unable to deal with cases of multi-modality where intensities are inherently different and where mutual information-based methods may be more appropriate.

Sum-of-Squared-Differences

One of the most intuitive and least resource-intensive approaches is the sum of differences and its variants. Pixels are being compared in two images, one pixel at a time, and their (potentially squared) grey-level difference are calculated. A sum over all pixel-wise differences is accumulated or averaged over, which gives a measure that is based on the sum-of-squared-differences (SSD). Mean-of-squared-differences (MSD) is merely the case where the differences are averaged over, rather than summed up. Other variants include the case where absolute differences are not raised to the power of two (squared). This method is usually powerful if the two images which are being compared are also closely aligned and their intensity values are relatively continuous and low in contrast. In cases where images are resistant enough to noise, MSD/SSD will tolerate a low level of locally-situated difference, while contrariwise, MI and NMI

properly handle greater dispersion of pixels in some localised region [77].

Mutual Information (MI)

Viola [120] developed a method[4] of measuring similarity between two images by repeatedly comparing histograms of pairs of images [80, 62]. When measuring mutual information, one computes *informational overlap* across images. The basic idea is that if two images are properly aligned, image values in one image can be used to predict to some extent image values in the second image. This means that the joint histogram for the two images will tend to contain sharp peaks. Under the complementary case of mis-registration, we will be unable to predict values in one image from values in the other image, which corresponds to a joint histogram without sharp peaks.

Let there be two images A and B. By defining a joint information (or entropy) to be $H(A, B)$ and the information contained in a single histogram A to be $H(A)$, then the mutual information (MI) is given by: $H(A) + H(B) - H(A, B)$. There are variants thereof [92], but the prime idea is that joint information is subtracted from the sum of information present in the two individual images.

[4]The discovery of mutual information is also attributed to Maes, yet some argue that the work was sparked by Viola in the mid-nineties. It is now accepted that Collignon *et al*, and Viola and Wells, came with the idea of mutual information for registration independently and at about the same time. Collignon, Maes *et al* published in 1995, and Viola also published in 1995.

One important strength of MI-based measures is that they are able to handle images that are obtained by different imaging techniques, hence unlike SSD, are not restricted to images of a single modality, but can perform multimodal registration.

Normalised Mutual Information (NMI)

Studholme [104] and Maes [65] suggested that normalisation should be applied to mutual information. Several steps are involved in this normalisation process[5]. The main difference is that the expression used for MI is significantly extended and divided by a normalisation term. The method is predominantly used in non-rigid registration as it is generic, adaptable to new data, and yields better results.

Correlation Ratio

Another method for measuring similarity makes use of the correlation ratio, the principle of which goes back over half a century ago [53]. The correlation ratio is used to solve a broad range of problems, but in this case it estimates a measure of the relationship between the dispersion in an image set and the dispersion across a larger sample. It takes into consideration a small set of images – however they may be defined – and

[5]There is an additional distinction between symmetric and asymmetric normalised mutual information, but rationale for this requires the full technical recipe. The dissertation at `http://www.lans.ece.utexas.edu/~strehl/diss/node107.html` summarises the way in which NMI is evaluated.

compares that to a larger set of images. The correlation ratio takes values in the range of 0 to 1.

Feature-based methods

Methods that are based on mutual information or MSD have become popular for image registration, but underlying image features is rarely accounted for. Instead of considering the structure of images as complete entities, it is possible to assimilate *portions* of image features, such as lines, intersections of lines, boundaries, corners, curves, and particular points of greater significance [99, 132]. Image intensity need not be the default feature.

Feature-based methods assume that intensity alone is not sufficient for actual *features* to be matched, so one approach to take is to use feature images, which may also contain the original intensity image. Extraction of features from the intensity images can be done separately (e.g. using image segmentation) or obtained using some assumptions about the imaged object. In some cases, features can be identified by considering other modalities where features in the images are more pronounced.

To use a practical example, an algorithm that pays attention to relevant geometric primitives inside images can calculate distances. It may use these distances to form a measure of dissimilarity while not accounting for raw image intensities at all, despite the fact that there is often a relationship between intensity and underlying features.

There are drawbacks to this approach because point- or curve-based NRR techniques might lack automation. They sometimes require user input in order to be robust and reliable. They can also can be negatively affected by poor detection and selection of edges, or be misguided by noise, to which they are arguably less resistance than area-based registration methods.

2.2.4 Optimisation

The nature of the image registration problem means that the search space for the optimisation algorithm is typically very high dimensional. In many cases, a general optimiser is used to solve the problem. In other cases, a more problem-specific optimiser is devised. In the case of free-form deformations, for example, the optimisation is simple because it is possible to deal with just one parameter at a time. In the case of fluid registration, the derivative of the objective function, which is indicative of force, is used to is used to drive the fluid flow. It is apparent that the warp representation, whether it is based on fluid, FFD, or even basis functions, affects the choice of optimisation method. A standard optimiser – even though it might be effective – simply does not take advantage of the nature of the warp representation and the data at hand.

Suggestions have been made over the years with regard to the issue of speeding up similarity measures and NRR in general [108]. The above

measures depend heavily, from an efficiency point-of-view, on the dimensions of an image. A multi-resolution approach [107] has plenty to offer in this case. For example:

- It can be used to speed up the entire process. Blurring or averaging, followed by re-sampling or sub-sampling, allows images of smaller size to be manipulated and complexity to be quadratically lessened. As the similarity measures are proportional to the images size, far better performance can be achieved with transition from coarse to finer resolution. Pluim *et al.* [81] studied the effects that this approach will have on the measurement of similarity.
- A multi-resolution has other advantages, such as its ability to make the optimiser avoid local minima.

2.2.5 Pairwise versus Groupwise Registration

Originally, NRR was based on comparison between pair of images. When a group of images is presented, it is not clear whether to treat separate pairs from this group separately and repeatedly or to make use of the entire group.

A distinction is made between two approaches to tackling the NRR problem. Some take the approach wherein one image from a set is chosen as a reference (or template) and the remainder of the set is transformed to

fit that reference [92, 91, 34]. It is a repeated pairwise approach and it means that, at the very end, all images will be assimilated to that particular reference, which was arbitrarily and possibly ill chosen. The other approach is based on the idea that the entire set of images should be transformed to minimise a groupwise objective function [22, 21, 68].

Debates over the validity (or lack thereof) in the pairwise approach are of great relevance to the work presented hereafter.

In repeated pairwise registration, there are various choices we can make as regards a reference image. However, the results of the registration tend to be dependent on this choice. For example, choosing a reference image from the set of original images will in general bias the results, with different choices producing different registrations. This motivates the need to evaluate the results of NRR algorithms [16, 93, 66].

Another issue is when non-deterministic choices are made in the registration algorithm, which can also lead to choice-dependent results.

Because of these problems, the work described in this thesis concentrates on a fully groupwise approach, with the aim of finding an approach that is not subject to these ambiguities and biases.

2.3 Assessment of NRR

A single NRR algorithm yields a particular solution to a problem, and different NRR algorithms tend to produce different results. It is hence necessary to have some way of quantifying the results of registration, to give us some standard by which to compare the performance of various algorithms, or to guide improvement of a single algorithm.

Two main approaches to assessing the accuracy of NRR algorithms have been described in the literature – one based on the recovery of known deformation fields, the other based on measuring the overlap of ground-truth annotations after registration. Both approaches are valid, but neither is easy to apply routinely, and both are better suited to off-line evaluation of algorithms, rather than *in-line* evaluation of the results of NRR in practical applications [43, 66, 103].

2.3.1 Recovery of Deformation Fields

One way to test the performance of a registration algorithm is to take some image data and then apply known artificial warps to it [87, 96]. This means that the ground-truth correspondence is known. This approach can be used to compare the performance of different NRR algorithms but, since it relies on the creation of artificial test data, cannot be applied in-line. Also, the validity of the approach depends on the ability to construct

artificial deformations which mimic the variability found in real images of a given type, which is difficult to guarantee.

2.3.2 Overlap-Based Assessment

An alternative approach is based on measuring the alignment [46], or overlap [46, 87] of anatomical structures annotated by an expert, or obtained as a result of (semi-)automated segmentation. This has the disadvantage that manual annotation is expensive to obtain and prone to subjective error, whilst reliable automated or semi-automated segmentation is extremely difficult to achieve – indeed if it was available it would often obviate the need for NRR.

Let us suppose that such annotation is available, so that we are provided with pixel-by-pixel binary label information for each image. For example, for an image of a brain, the set of labels could be tissue-type labels such as CSF, white matter, grey matter, or at a finer level of detail corresponding to individual structures. The Tanimoto overlap between a pair of binary label images A and B is considered here (see Figure 2.4). Tanimoto overlap [7] is the ratio between the intersection of A and B, and the union of the two. This can also be written $O_p = \frac{N(A \cap B)}{N(A \cup B}$ which deals naturally with cases where applying the deformation field to a label image results in label values between 0 and 1.

This idea can be generalised to the case of a group of images with multi-

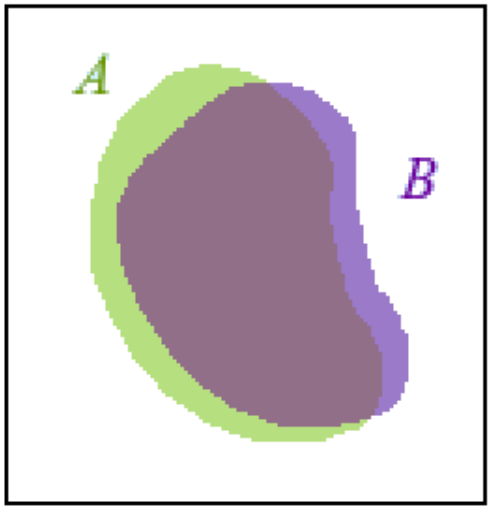

Figure 2.4: Two binary labels, A and B, overlap one another

valued tissue labels for each voxel. Each label for a given image is represented using a binary image but, after warping and interpolation into a common reference frame, based on the results of NRR, a set of fuzzy label images is obtained. These are then combined in a generalised overlap score [16] which provides a single figure of merit aggregated over all labels and all images in the set:

$$\mathcal{O} = \frac{\sum_{\text{pairs},k} \sum_{\text{labels},l} \alpha_l \sum_{\text{voxels},i} MIN(A_{kli}, B_{kli})}{\sum_{\text{pairs},k} \sum_{\text{labels},l} \alpha_l \sum_{\text{voxels},i} MAX(A_{kli}, B_{kli})} \tag{2.1}$$

where i indexes voxels in the registered images, l indexes the labels and k indexes image pairs (all permutations are considered). A_{kli} and B_{kli} represent voxel label values for a pair of registered images and are in

the range $[0, 1]$. This generalised overlap measures the consistency with which each set of labels partitions the image volume. The standard error in $\mathcal{O}$ can be estimated in the normal way from the standard deviation of the pairwise overlaps.

The parameter α_l affects the relative weighting of different labels. With $\alpha_l = 1$, label contributions are implicitly volume-weighted with respect to one another. This means that large structures contribute more to the overall measure. One can also consider the case where α_l weights labels by the inverse of their volume (which makes the relative weighting of different labels equal), where α_l weights labels by the inverse of their volume squared (which gives regions of smaller volume higher weighting), and where α_l weights labels by their complexity, which is defined as the mean absolute voxel intensity gradient over the labelled region.

An overlap score based on a generalisation of the popular Dice Similarity Coefficient (DSC) would also be possible but, since DSC is related monotonically to the Tanimoto Coefficient (TC) by DSC = 2TC/(TC+1) [98] it need not be considered further.

2.3.3 Motivation for Ground-Truth-Free Assessment

The search for an assessment method that uses the image data alone is motivated by numerous factors:

1. The need for an assessment that is unaffected by human judgement such as annotation, where there can be considerable inter and intra subject variation.

2. The need to process large amounts of data, either in 2- or 3-D. Without automation, the process is hindered by human capacity.

3. The resources allotted to obtaining ground truth. Different types of data are inherently different, so methods of extracting ground truth may vary.

The method sought would potentially be resource hungry. The subsequent chapters show that a brute-force approach can be used as a substitute to an approach that uses ground truth.

The need for ground-truth-free methods of evaluation has been recognised in other areas. The STAPLE algorithm [126] from Warfield *et al*. addresses the problem of evaluating complete segmentations of an object. The method uses an E-M algorithm to estimate the most likely ground truth, given the set of segmentations.

2.4 Summary

In this chapter, we have introduced the basic concepts of non-rigid registration, and discussed the issues involved in constructing and assessing

registration algorithms. In the next chapter, we describe statistical modelling techniques which can be applied to image data once the results of registration have been obtained.

Chapter 3

Active Appearance Models

"If you optimize everything, you will always be unhappy."

– *Donald Knuth.*

THIS background chapter covers models that are capable of generating images of a particular class, e.g. MR images of the brain. It concentrates on statistical models of shape and appearance, where models are learnt from sets of examples images. The motivation of this approach is presented, followed by technical details of model construction. These models can be used in a variety of ways, which includes their ability to synthesise images and interpret images through fitting.

3.1 Motivation

In image analysis, segmentation is a key step. Segmentation identifies different regions of a particular image, so understanding of structure can be gained. There are two broad approaches to solving the segmentation problem. The first is the bottom-up approach, which involves grouping of pertinent parts of an image. The grouping then enables one to tell apart different regions and understand the relationship between them. The second approach to segmentation is the top-down approach. It relies on a model which represents some prior knowledge (or assumptions) about the image being interpreted. The model organises image evidence to provide an 'explanation' of the image, from which the statistics can be obtained directly.

There are advantages and disadvantages to each approach. The bottom-up approach makes few assumptions, but it can be difficult to achieve reliable results for complex images. The top-down approach uses prior knowledge that includes reliable segmentation, but cannot deal with situations where the model does not apply.

This chapter takes a look at the second approach to image segmentation. It describes models of shape and appearance, particularly the approach of Cootes and Edwards [20, 32, 31, 19], who introduced models that capture variation in both shape and texture (in the graphics sense). These models have been used extensively in medical image analysis in appli-

cations ranging from brain morphometry to cardiac time-series analysis [34, 91, 101], as well as their extensive use in non-medical imaging applications such as face recognition.

The approach is an example of 'interpretation by synthesis' – it uses a 'generative' model that is capable of synthesising examples of the class of images to be interpreted, and image interpretation of a given image proceeds by matching the synthesised image of an object with the actual example of the object within the image being interpreted.

To use an example, Figure 3.1 illustrates this approach. The model M is positioned in the image close enough to its target, which is the spine. The model contains knowledge about ways it is allowed to deform. Model parameters are repeatedly changed until the model (ideally) overlaps the target images. Then, by extracting model parameters, knowledge can be gained about the image itself. This method has proven useful in various areas, including industrial inspection, motion analysis [129], face recognition [24, 61], and medical image understanding [102].

The remainder of this chapter introduces statistical shape models, explains how they can be extended to full appearance models, and then describes their use in image interpretation.

Figure 3.1: A target spine image T, overlaid by a high-level representation (the model M), which searches for an improved fit in the target image by transforming itself.

3.2 Shape Models

Given a collection of images with common properties, it is possible to model visual form (or *shape*) of the objects therein. A model is a simplified representation of structures that vary across the images, and can be built in a way that makes it independent from subtle changes in view-point, object position, and size [4]. The models used here are statistical, using the notion of a probability density to describe shape variation.

The model learns about the variation of shape from a collection of shape examples (the training set).

Suppose we have such a set of shapes. The first issue is that we need to remove variation such as position, orientation, and scale of the entire shape which has nothing to do with the actual variation of shape itself. So, we are treating a shape as being the same shape, whatever position

it appears in an image, whatever size it happens to be, or whatever its orientation. This is the process of shape alignment.

The second stage then involves constructing a way to represent the variation of the set of aligned shapes, and to do this we need a useful way to represent the shapes themselves.

3.2.1 Representing Shapes

If a human annotates the shape of an object within an image, this is typically done by them simply tracing the outline of the object. However, such a continuous outline is not directly suitable for analysis purposes.

The simplest way to represent such an outline in a compact fashion is by some set of points (called landmarks or shape points). The entire shape is then obtained from the positions of these landmarks by interpolation. The simplest case is linear interpolation, where the landmarks are the vertexes of a polygon, but more complicated spline interpolants are possible if a smoother outline is required.

This method can describe a single shape, but what we require is a representation which describes a set of shapes within a single common framework. This is usually achieved by identifying corresponding landmarks across a set of shapes (see Figure 3.2 on page 59). For example, if we are modelling the shapes of biological objects, landmarks can be placed at anatomically relevant points, which can be reliably located in any

example of the object. Further shape points can then be created semi-automatically along edges in the images, to produce the final dense polygonal representation.

Mathematically, the position of each landmark point within each image is encoded in terms of the Cartesian coordinates of each point. The entire set of landmark points within a given shape is then represented by a single shape vector $\mathbf{S}_i$, formed by concatenating the coordinates of the individual landmarks

$$(x_1, y_1, x_2, y_2, ..., x_n, y_n) \Rightarrow \mathbf{S}_i. \tag{3.1}$$

3.2.2 Aligning Shapes

We now have our set of shapes in a common representation. However, we first have to remove the degrees of freedom corresponding to position, orientation, and scale, as mentioned above. The method most commonly used is Procrustes analysis. The generalised Procrustes procedure was developed in 1991. Procrustes analysis translates, rotates and scales each shape, so as to minimise the sum of squared distances between corresponding landmarks. Having aligned all the shapes, the mean shape is then obtained by simply averaging the position of each corresponding landmark across the whole set of shapes.

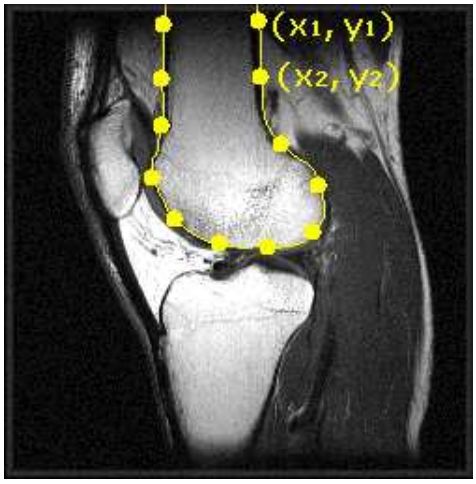

Figure 3.2: Landmark identification and mark-up in medical images

3.2.3 Obtaining Modes of Variance

We now have our mean shape, and the set of training shapes within a common representation.

The variation of shapes is typically described in terms of variation about the mean shape. Specifically, how the landmark points vary about their mean position, and how the movements of individual landmark points are correlated across a shape.

In order to learn how the positions of landmark points vary, one can perform Principal Component Analysis (PCA). In PCA, eigen analysis is used to discover the modes of variation that best describe the variation seen across the set of shapes. These shapes, which lie in a common frame of reference and are each represented by a vector, can be thought of as points in a high-dimensional space (see Figure 3.3). What is required is a

set of vectors/axes in this space that best describe the subspace in which the shape data lies. PCA provides this by extracting the eigenvectors and eigenvalues of the covariance matrix of the data set.

These eigenvectors form an orthonormal set, whose importance is encoded in terms of the relative size of the associated eigenvalue. Axes aligned with these eigenvectors then provide a new coordinate system which spans the data subspace. Shapes are now described in terms of their components relative to these PCA axes, rather than the Cartesian coordinate axes that were used initially, and these PCA axes capture the correlations between the motions of landmarks.

All correlations are not equally important, and the relative importance of the PCA components is encoded by the relative sizes of their eigenvalues.

This means that we are not required to use the entire set of PCA components, but can instead choose a restricted set, that keeps only those components that are the most significant. The process is hence lossy. That loss is being controlled, however, in the sense that one can choose the minimal amount of variation which must be accounted for[1]. What this boils down to is the building of a model that is smaller in size (because we do not retain all PCA components), yet one which also encapsulates the significant modes of variation of the data, and the significant correlations between landmark positions.

[1]A sensible choice might be, for example, 98% of the observed variation, which means that 2% of the variation is not accounted for. In practice, that 2% of the overall variation is usually the least informative and it is possibly made up from noise and error. Annulling this effect is, among other things, what PCA is intended to accomplish.

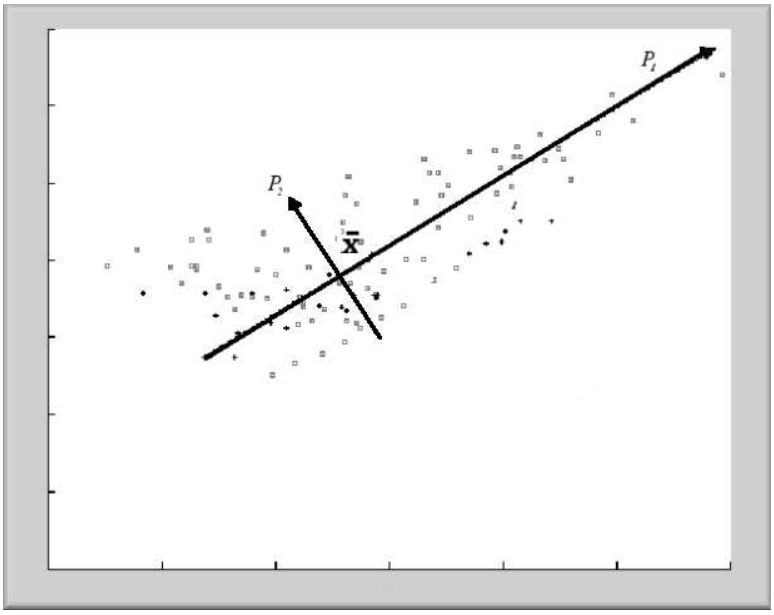

Figure 3.3: The two principal components in a 2-D data scatter is indicated by the two arrows

We now have our mean shape and restricted set of PCA axes that provide a concise description of the training set of shapes.

However, this model can also be used to generate new shapes. Let such a new shape be $\mathbf{x}$, generated from a set of shape parameters $\mathbf{b}_s$:

$$\mathbf{x} = \overline{\mathbf{x}} + \mathbf{P}_s \mathbf{b}_s. \tag{3.2}$$

The matrix $\mathbf{P}_s$ contains the eigenvectors of the covariance matrix of the training data.

The generated shapes can be constrained to be similar to those seen in the training set, by constraining the allowed shape parameters, b_s, to be similar to those extracted for the training shapes. Typically, the distribution of training set shape parameters is modelled by a multivariate Gaussian pdf, and new shapes are generated by sampling from this pdf.

Figure 4.2 on page 78 shows an example where model parameters are altered. The effect of varying parameters that correspond to first of second modes of variation is shown using hand data.

3.2.4 The Correspondence Problem

It is important to note that if our only task was representing a set of shapes to some required degree of accuracy, we could select a dense set of landmarks on each shape, with no correspondence between landmarks on different shapes. This however would perform poorly if we used these landmarks to try and describe the variation between shapes. The simplest way of obtaining meaningful corresponding landmarks across a set of shapes landmarks is for a human expert to annotate the training shapes with the positions of equivalent points using some computerised special-purpose tools. In recent years, alternatives which are automatic showed great promise [25] and they were also extended to 3-D [27]. A subsequent background chapter on MDL shape models is dedicated purely to that one strand of work, so we will not discuss this further here.

3.3 Appearance Models

In the previous section, we described how to build statistical models that both describe a training set of shapes, as well as encapsulating information about the variation across the set of shapes.

It is possible to extend this method, and construct models that encapsulates not just the variation of the shape of objects in images, but also the variation in the appearance of the object itself.

Appearance models were developed by Edwards *et al*. Their greatest contribution, advantage, and essence lie within the fact that they incorporate textural information rather than shape alone. Texture is a made out of meshes of grey-level pixel intensities. Incorporation of full colour is possible as well [101]. Colour can be simply thought of as an extension of the single grey-scale band. It can be divided up into bands using the most common separability: red, green, and blue components [2].

A shape model can be thought of as providing very limited information about the appearance of an object within an image, in that it describes the shape of an object, where it is implicitly understood that the shape of an object corresponds to strong edges in an image.

Appearance models describe not only the shape of an object, but the image intensities within the outline of the object as well. In the following

[2]There are different possible colour schemes [78], but they need not have any affect on principles of sampling intensities.

subsections, three steps are discussed in turn: modelling shape, modelling intensity, and combined models.

3.3.1 Modelling Shape

The first step is building of shape models, which use a finite number of modes for representation. Shape models that are built from the outlines of objects in images enable those images to be brought into a state of alignment. We can warp the shapes within an image to match the mean shape. By interpolation, we can extend this warp to the entire interior of the object. This means that corresponding parts of shapes in those images will be easy to identify and then use, e.g. in order to sample intensities.

3.3.2 Modelling Intensity

To model texture, differences in shapes are removed by morphing each training image to the mean shape[3]. A shape-free texture patch can then be estimated from the image by sampling on a regular grid and forming a vector g. Statistical analysis proceeds as for shapes and it results in the following linear expression for texture

[3]Warps can be performed using a strategy borrowed from graphics. In all experiments described in this thesis this was achieved by a triangulated mesh generated from the landmark points and barycentric coordinates to mesh the intersections put in vector g (Equation 3.3).

$$\mathbf{g} = \overline{\mathbf{g}} + \mathbf{P}_g \mathbf{b}_g. \tag{3.3}$$

$\mathbf{g}$ is the intensity vector. $\mathbf{P_g}$ contains the eigenvectors of the covariance matrix of the training data and $\mathbf{b}_g$ controls the intensity. The process is hardly different from dimensionality reduction in the case of shape.

Generally, we wish to distinguish between the meaningful shape/texture variation of the objects under consideration, and the apparent variation in shape/texture that is due to the positioning of the object within the image (the pose of the imaged object). In this case, the appearance model is generated from an (affinely) aligned set of images, just as was the case for shape models considered earlier. Point positions $\mathbf{x}_{im}$ in the original image frame are then obtained by applying the relevant pose transformation $T_{\mathbf{t}}(\cdot)$:

$$\mathbf{x}_{im} = T_{\mathbf{t}}(\mathbf{x}_{model}) \tag{3.4}$$

where $\mathbf{x}_{model}$ are the points in the model frame, and $\mathbf{t}$ are the pose parameters. For example, in 2-D, $T_{\mathbf{t}}$ could be a similarity transform with four parameters describing the translation, rotation, and scale of the object.

In an analogous manner, the image can also be normalised with respect to the mean image intensities and image variance,

$$\mathbf{g}_{im} = T_{\mathbf{u}}(\mathbf{g}_{model}), \tag{3.5}$$

where T_{u} consists of a shift and scaling of the image intensities. For further implementation details see [20, 32].

As noted above, a meaningful, dense, groupwise correspondence is required before an appearance model can be built. NRR provides a natural method of obtaining such a correspondence, as noted by Frangi and Rueckert [34, 91]. It is this link that forms the basis of the new approach to NRR evaluation.

The link between registration and modelling is further exploited in the Minimum Description Length (MDL) [119] approach to groupwise NRR, where modelling becomes an integral part of the registration process. This is one of the registration strategies which is discussed in later chapters.

3.3.3 Combined Models

The models in Equation 3.2 and Equation 3.3 have a linear form, so they are quite compact. This is a highly desirable property which makes the models flexible and manageable.

However, at the moment the two components of the model, the shape x and the shape-free texture g are independent. In real images, shape and texture are not necessarily independent. The easiest example is to think of an image of an individual face. When the person changes expression,

the shape of the face obviously changes. But the texture (i.e. the positions of highlights and shadows) obviously changes too, in a way that is correlated with the shape change. Hence it is desirable to merge the individual shape and texture models, to obtain a new model that is aware of both types of variation. This combined model can then also incorporate any correlations between shape and texture.

The parameters $\mathbf{b}_s$ and $\mathbf{b}_g$ are aggregated to form a single column vector

$$\begin{Bmatrix} \mathbf{b}_s \\ \mathbf{b}_g \end{Bmatrix}. \tag{3.6}$$

The new vector is a simple concatenation of the two. However, since the values of intensity and shape can be very different in magnitude, weighting is needed. Such weighting brings equilibrium, under which both shape and intensity maintain a sufficiently-noticeable effect and impact on the model they jointly build. A weighing matrix resolves the problem introduced here and it is, by convention, named $\mathbf{W}_s$[4]. With weighing in place, aggregation takes the form

$$\begin{Bmatrix} \mathbf{W_s b}_s \\ \mathbf{b}_g \end{Bmatrix} \tag{3.7}$$

[4]The letter s stands to *shape,* as by default this matrix scales the shape parameters only. It gives logically equivalent results to these of applying the factor $\mathbf{W}_g = \frac{1}{\mathbf{W}_s}$ to intensities.

where $\mathbf{W}_s$ is set to minimise inconsistencies due to scale. By applying another PCA step to the aggregated data, the following combined model is obtained

$$\begin{aligned} \mathbf{x}_i &= \bar{\mathbf{x}} + \mathbf{Q}_s \mathbf{c}_i \\ \mathbf{g}_i &= \bar{\mathbf{g}} + \mathbf{Q}_g \mathbf{c}_i \end{aligned} . \tag{3.8}$$

The appearance (shape and texture) is now purely controlled by the new set of parameters, $\mathbf{c}$. There is no need to choose values for two 'families' of distinct parameters. This combined model reaps the benefits of the dimensionality reduction performed, which is based on shape as well appearance. This means that this new model encompasses all the variation learned and the correlation between these two distinct components. Since PCA was applied, the number n of parameters $\mathbf{c}_i$ is expected to be smaller than the number of parameters in $\mathbf{b}_s$ and $\mathbf{b}_g$ put together.

The effect of varying different elements of a combined model $\mathbf{c}$ for a model built from a set of 2-D MR brain images is shown in Figure 3.4. The number of modes (columns) in $\mathbf{Q}_s$ and $\mathbf{Q}_g$ is one less than the number of images. In practice, it is often possible to approximate images well, using fewer modes m.

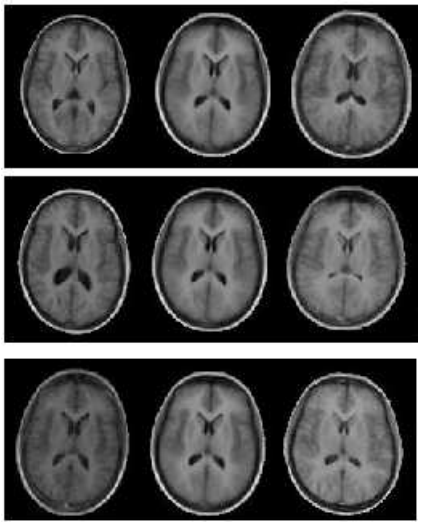

Figure 3.4: The effect of varying the first (top row), second, and third parameter of a brain appearance model by ± 2.5 standard deviations

3.4 Active Appearance Models

Given an appearance model, this model can be further extended and be used for image interpretation. One can take the model and select a set of parameter values which make the model resemble a given image. It is a case of matching to an image for analysis purposes and it is treated as an optimisation problem. This process is sometimes known and referred to as "model fitting" because a model gets manipulated until it best *fits* an image. Matching (or fitting) is made possible thanks to a process characterised by model *training*. It is the first step which involves extending an appearance model and having it equipped with additional knowledge (learning how to fit the model). The second step involves searching for matches in unseen images.

3.4.1 Learning to Fit the Model

This first problem can be solved by learning how the parameters $\mathbf{c}_i$ (see Equation 3.8) affect the model and how they should be changed in order for the model to align with the image to be interpreted. Each parameter in $\mathbf{c}_i$ has an effect in model synthesis (both shape and intensities vary simultaneously). By changing the value of each such parameter and learning the difference between the model and the target image (using pixel-based comparison), a index can be built which indicates how model parameters should be varied to better fit the target. This type of index indicates which parameters should be changed and, if so, in what way and to what degree.

More formally, the algorithm can be described as follows. For the model parameters $\mathbf{c}_i$, where $1 < i < n$, a parameter change $\delta\mathbf{c}$ (one parameter or more can be changed simultaneously) is applied to generate new model synthesis consisting shape and texture. This process is repeated for each mode of displacement where both shape and intensity are changed, but only intensity difference is learned. Sum-of-squares of the pixel differences[5] is then used

In [19], a similar formulation is used. Shape transformation parameters, $\mathbf{t}$, are used to define point position in $\mathbf{X}$, which is the image frame of the model. The pixels in the region of this image, $\mathbf{g}_{im}$, can then be projected

[5] A simple measure of difference is used here although this need not necessarily be the case.

into the texture frame $\mathbf{g} = T_u^{-1}(\mathbf{g}_{im})$ and the model texture is $\mathbf{g_m} = \bar{\mathbf{g}} + Q_g\mathbf{c}$. The image difference is $\mathbf{r}(\mathbf{p}) = g - g_{im}$. From this, the image residuals can be derived. This is made possible after a Taylor expansion that yields a convenient term to optimise over. This term is referred to as $\delta\mathbf{I}$ from this point onwards.

With this measure of intensity difference, the relation between the parameter change and this difference can be expressed conveniently. That information, which is merely a correlation, can be learned by using a pseudo-target image, such as the model in its standard form (the mean). It can be used as a basis for a comparison which facilitates learning about the model displacements and their corresponding effects in image space.

This quantitative measure of difference will approximate "goodness" of a parameter change (as indicated when calculating SSD or MSD), but not more *localised* effects which such a change has on the given image. This means that it will not be obvious what parts in the two entities (model and target) remain similar and which ones do not.

To address the need for localised difference measures, the parameter change, $\delta\mathbf{c}$, which is applied to the collection of parameters $\mathbf{c}_i$, will be accompanied by a high-dimensional representation of intensity differences in the image. This correlation can be made accessible through an index whose size is proportional to the image size. The relationship is dictated by the following:

$$\delta \mathbf{c} = \mathbf{A}\delta\mathbf{I} \tag{3.9}$$

where $\mathbf{A}$ is a matrix[6] that encapsulates the change in intensities due to the parameter/s change δc. This matrix is correspondent to an n-dimensional vector. This vector expresses the change which was discovered 'off-line', i.e. prior to an search in an image. It linearly defines (in a potentially high-dimensional space) the linear relationship between change to the parameters and change to the intensities (the *difference image*). It can be used to choose directions of change directly when performing a search and thereby avoid re-computation.

3.4.2 Model Fitting

The second stage in active appearance models involves use of the model obtained, along with the correlations learned for that model, in order to do the fitting [97]. It is possible to perform an image search which is driven by the observed difference between the model and a given target image. Fitting is done by changing the values of model parameters. The state of the active model is defined by parameter values which, assuming the model is fitted successfully, contain information about the target image.

The model, as illustrated in Figure 3.5 and Figure 3.1 on page 56, is ini-

[6]The matrix $\mathbf{A}$ can be obtained using linear regression.

tially placed somewhere inside the image frame, with reasonable proximity to its target. The reason why good proximity (initialisation) is essential is that significantly large displacements are rarely learned off-line. The model is quite meaningless unless there is at least partial overlap or commonality to drive the optimiser in the right direction.

The algorithm which is used to perform the search adopts the following general form:

- Place the appearance model **M** somewhere in the image, preferably at the centre where the target of interest (to be denoted by **I**) is likely to lie[7].
- With the appearance model in its current state and the static target, perform the following:
  - ✧ Calculate the differences between the model and the target. This can be done by synthesising **M** and calculating δ**I = M - I**.
  - ✧ Save the new state of the appearance model if the difference has been lowered, i.e. similarity is being approached.
  - ✧ Adjust the value of Euclidean parameters, potentially with inclusion of a scaling coefficient $k = 1.5, 0.5, 0.25$ and so forth. This often achieves good results, although it is a heuristics-driven technique.

[7]Advanced knowledge about the problem is highly conductive at this stage. Otherwise, a bottom-up image analysis is a must.

Figure 3.5: Model and target fitting.

- ✧ Using the correlations learned off-line[8], set new values for the parameters $\mathbf{c}_i$ of $\mathbf{M}$ where $\delta\mathbf{c}$ is a function of k (see above).
- ✧ Compute the new difference measures between the model and the target (as previously).

- Iterate while no state of convergence has been reached and improvements are still observed at times.

The technique of matching an appearance model to a target image can be depicted using a large sequence of images (an animation) resembling

[8]If these correlations are not available, guessing would be an alternative. It is important, however, to learn from the experience gained during this independent run of the program or else the optimisation would behave senselessly and lead to improvements being identified very slowly. General optimisers are assumed to make a good judgment as such.

one which is shown in Figure 3.5. Only a few dozens of iterations are required in order to get good matching. This depends on the algorithm and the scale of the problem.

3.5 Summary

The chapter covered and demonstrated the use of combined models of shape and appearance. Their formation process, which relies on the ability to identify and improve the correspondence, was a key component of this. The next chapter concentrates on work which aims to improve these correspondences automatically and thereby improve the quality of models that are built.

Chapter 4

MDL Shape Models

"All generalizations, with the possible exception of this one, are false."

– *Kurt Gödel.*

In the previous chapter, generative models of shape and intensity were described. Construction of such models depends on the ability to establish correspondence, which is a central factor that affects the quality of models. Over the past few years there has been considerable research on establishing correspondence between a set of *shapes* automatically. This chapter provides a succinct summary of this work.

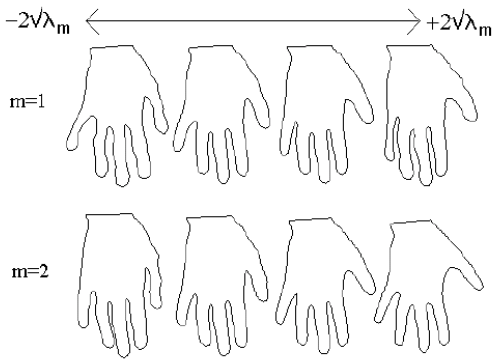

Figure 4.1: The first two modes of variation of a model built from the equally-spaced annotated training set. Figure from Rhodri Davies [25].

4.1 Overview

Figure 4.1 shows that picking landmark points arbitrarily (or just picking points that are equally spaced) leads to a model which is poorer than one where careful manual selection of correspondences was involved (as shown in Figure 4.2).

By comparing the two figures, it should become evident that if one selects the wrong correspondences, one will obtain poor models. The model of the hand should ideally be capable of encapsulating something which does not result in distortions of a natural form of a hand, as illustrated by the first mode of variation in Figure 4.1 (notice how thickness of the fingers varies).

This chapter addresses the need to find good correspondences and explores some previous work. In particular it considers an approach where

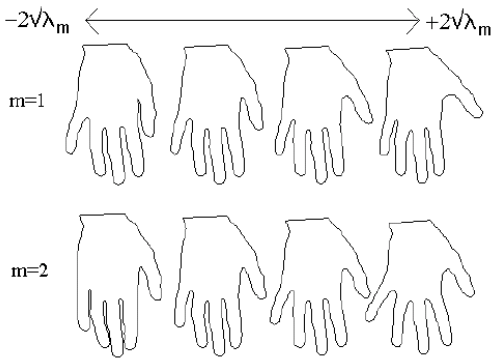

Figure 4.2: The first two modes of variation of a model built from manually-annotated hand data. The figure shows the effect of independently varying the first two modes of variation. Figure from Rhodri Davies [25].

establishing correspondence is posed as an optimisation problem, where the aim is to select the correspondences between shapes in the training set that result in the best model.

This correspondence-by-optimisation approach requires 3 things:

1. Method of measuring model quality;
2. Method of manipulating correspondence;
3. Method of optimising.

Each one will be dealt with in turn.

4.2 Measuring Model Quality

Correspondence-by-optimisation requires an objective function that provides a measure of model quality for a given set of correspondences. In early work, Kotcheff and Taylor [54] proposed the use of the determinant of the shape covariance matrix as a measure of model quality, arguing that a good correspondence should result in a model with a small distribution in shape space. They showed encouraging results but did not provide a sound theoretical justification for the approach.

Davies *et al.* [26] built on this work, providing a Minimum Description Length (MDL) objective function. Their basic idea was that a good model ought to allow a simple description of the training set – in an information theoretic sense. The MDL objective function estimates the information (in bits) required to encode the set of shapes using the model.

For one-dimensional centered Gaussian models, the shapes and models were encoded using an expression that considers the family of such Gaussian distributions and is computed by summing two parts; the description length for encoding value *and* accuracy of a parameter along with the description length for encoding the value itself using the model [25].

It is interesting to note that the MDL objective function includes a term that is closely related to Kotcheff and Taylor's determinant, and which often dominates. This means that the MDL objective function can often be approximated by the determinant-based objective function.

4.3 Methods of Manipulating Correspondence

Correspondences describe the relationship between shapes. In Chapter 3, it was shown how a shape model can be built once each shape has been annotated with a set of corresponding landmarks. To alter the correspondence between a set of shapes, one needs to find a method of moving the landmark points, and moving them in such a way so that they remain on the shape. The simplest way to achieve this is by building a parameterised representation of shape.

Taking as an example the simplest case of open-ended aligned shapes in 2-D, let the initial parameter be considered for each shape the fractional distance along the shape (arc-length parameterisation). This then gives a parameter t, which varies from 0 at one end of the shape to 1 at the other end.

For the i^{th} shape in our training set, at a given parameter value t, one gets the position in real space of that point on the shape, which is denoted by $S_i(t)$, where $S_i(\cdot)$ is now the shape function for the i^{th} shape.

Correspondence between these continuous shapes can now be defined by stating that points at the same parameter value correspond. That is:

$$S_i(t) \sim S_j(t). \tag{4.1}$$

This can now be used to manipulate correspondence. It is possible to

re-parameterise only the i^{th} shape, so that:

$$t \mapsto t' \doteq \phi_i(t) \tag{4.2}$$

where $\phi_i(\cdot)$ is then the re-parameterisation function for this shape. The shape function now also changes, although the shape itself of course stays the same. A new re-parameterised shape function $S'_i(\cdot)$ is defined where:

$$S'_i(t') = \S'_i(\phi_i(t)) \doteq S_i(t). \tag{4.3}$$

The physical point on the shape remains unchanged, but its parameter value changes from t to t', and the shape function changes accordingly. Since correspondence between shapes is defined for points of the same parameter value, this now also changes, from $S_j(t) \sim S_i(t)$, to $S_j(t) \sim S'_i(t) = S_i(\phi^{-1}(t))$.

The re-parameterisation function for the i^{th} shape, $\phi_i(t)$ must obey certain conditions if it is to be valid. The values of the new parameter t' must still lie between 0 and 1, each point on the shape should have a unique parameter value, and each parameter value should correspond to a single point on the shape. In mathematical terms, ϕ_i has to map the interval $[0, 1]$ onto itself in a one-to-one fashion, and is hence a monotonic function on this interval.

Figure 4.3 shows an example of the construction of a monotonic function

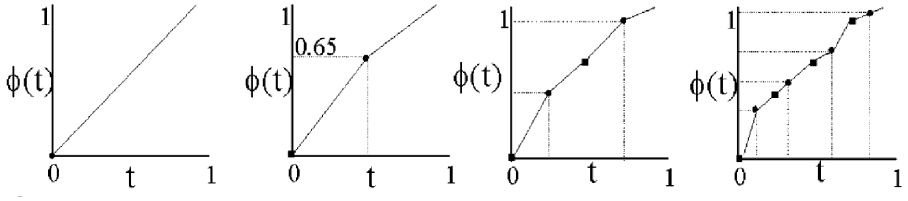

Figure 4.3: A piecewise-linear construction of a re-parameterisation function. The graphs at the top show the re-parameterisation function as the number of nodes increases. Figure from [26].

using piecewise-linear recursive re-parameterisation. Figure 4.4 shows the 2-D parameter space for shapes in 3-D with the topology of the sphere. The points indicate how points move under re-parameterisation.

Suitable re-parameterisation functions can also be built up in other ways other than by using linear interpolation. One approach is to use as the re-parameterisation function the cumulative distribution function of some normalised probability density function (pdf), which is then guaranteed to be monotonic and lie within the required range.

As in methods of kernel density estimation, a suitable pdf can be built by adding kernels, where the positions, widths, and amplitudes of the kernels are the parameters that control the re-parameterisation.

For example, Cauchy kernels [27] can be used, each kernel then having the 3 parameters of amplitude, width, and position. Other kernels that have been used include the von Mises distribution, cardioids, or the most obvious choice, Gaussian kernels. Such kernels then obviously lend themselves to a coarse-to-fine approach, where one starts with a small number

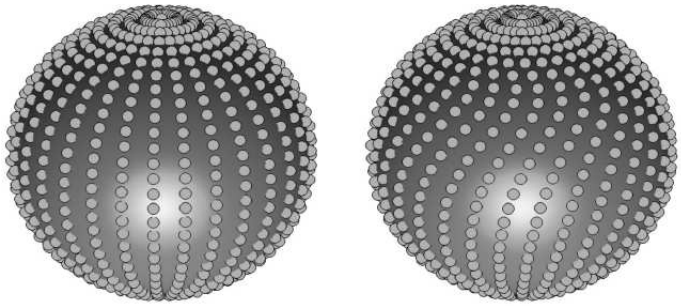

Figure 4.4: An example reparameterisation in a high-dimensional case. The sphere on the right illustrates the relocation of points after a kernel function was applied to reparameterise what is shown on the left. Figure from [25].

of widely-spaced kernels, and gradually add more finely-spaced kernels as the optimisation proceeds.

4.4 Optimisation

An optimisation strategy is needed in order to identify parameterisations that lead to better models. Various methods were explored in the attempt to reach good solutions and reach them quickly. A general-purpose optimiser seems like the most natural choice to make, but it does not necessarily produce good results. Such an optimiser, for example, can be faced with the task of handling very large sets of shapes. This slows down the optimisation and makes it less effective.

One alternative approach to optimisation is to handle shapes in isolation. A one-at-a-time optimiser will have a single shape and reparameterisa-

tion to handle while all other shapes are unchanged. This simplifies the problem and works well in practice.

Another key method that handles high complexity and scale is a multi-resolution approach. Coarser shapes of smaller size are derived from the original data. Gross estimates of a good solution can be found early on and then be propagated for use in optimisation over finer-scale data. Davies *et al.* found that it helped in cases where data was cluttered. Random selection of reparameterisation proved useful too.

The process of optimisation in the case above involved a number of steps:

1. Pick a shape
2. Pick a fixed breadth and width for the kernel function that generates the re-parameterisation
3. Optimise objective function with respect to kernel amplitude
4. Repeat (back to (1), until the results are satisfactory)

There were variants to this method as well, which proved worthwhile. For example, picking several kernels before moving on to next shape is a strategy that led to improvements.

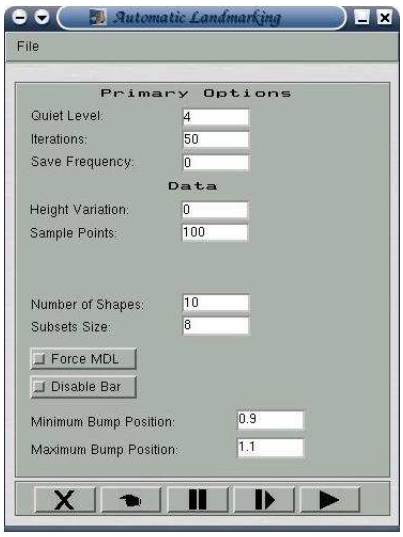

Figure 4.5: The graphical user interface for semi-automatic landmark selection.

4.5 Extensions

A series of experiments, which were primarily involved with improving optimisation and performance, are shown in this section. These are part of an important goal, which is to study strategies for optimisation that makes their application to images (both shape and intensity) feasible.

Experimental Framework and Data

In most of the experiments that follow, one particular data type is used to make up a set of similar shapes. It should be perceived as simplified test data that is used for the proof of concept. The data type is referred to as “brick and bump" (also called box-bumps by some authors), aptly named

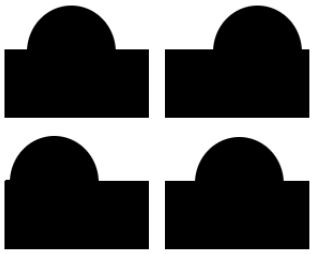

Figure 4.6: Four separate examples of "brick and bump" shapes. The shapes are closed curves. the inside is these curve is painted black only for illustration purposes.

so owing to the elements from which it is composed (Figure 4.6 contains some examples). The data is a rectangle with a semi circle (bump) on top of it. The only degree of freedom is the horizontal position of this bump. Figure 4.7 shows the principal modes of variation of a model built from improperly aligned brick and bump shapes. A newly-constructed front end (shown in Figure 4.5) handled the process of landmark identification. Like most applications that are used to carry out this work, the front end was constructed using Sun Microsystems' Java and Mathwork's MATLAB, under GNU/Linux. The new graphical user interface was used to carry out the majority of the experiments described in the remainder of this chapter, as well as in subsequent ones.

Consider Figure 4.7 which is intended to show not only the form of shape data which is being dealt with. It also depicts the results of using a poor point-to-point correspondence to construct a model of shape. Each point that is used to sample the shape is unique. The dots are connected using straight lines. Any point corresponds to another particular point in an-

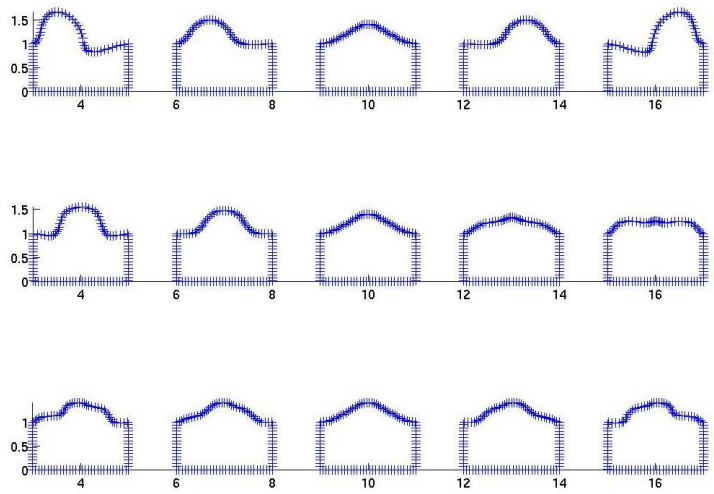

Figure 4.7: Unregistered bump-shaped synthetic data and its three principal modes of variation (± 2 standard deviations are shown). Each bump is composed of points, depicted as a plus sign, to make them distinguishable in the plots.

other shape. For a good correspondence, the set of shapes generated from the model built using this correspondence should appear similar to the shapes in the training set (Figure 4.6 on the previous page). However, it clear from Figure 4.7 that the generated shapes here do not look like the training set, containing dips and deformed bumps, rather than just a simple brick and bump.

Various experiments were then conducted to explore the possibilities of improved performance. Although comparative analysis was part of these experiments, detailed methods and results are omitted here.

Subsets

The idea of this approach is to speed up the algorithm by handling small sets. Instead of optimising with an entire large set at hand, subsets of it can be considered. Smaller portions of a larger problem are optimised in isolation and eventually this leads to quicker convergence. This division into subsets permits larger sets to be dealt with, going high enough to handle even hundreds of shapes—something which has thus far been impractical.

In practice the choice of shapes is made stochastically and changes repeatedly. By registering subsets, a globally-optimal shape model can be constructed.

Varying Optimiser Tolerance

As part of speed-up through modification, an adaptive precision approach was taken. The rate of convergence is changed as the process goes on and so is the speed/efficiency of the algorithm. There was much to be investigated in order to ensure the approach always results in gains. It was also worthwhile to see if the choice of tolerance can be made more preferable, based on empirical evidence. Experiments with varying values for tolerance showed that there was promise in an approach which seeks coarse and poorer solutions at the start and refines them further as it went along.

Experimental Results

The subset approach and variation of the optimiser's tolerance were both incorporated into the algorithm that identifies correspondence in shapes. Consider Figure 4.7 which shows a poor model built from the initial correspondence. It is clear that positions of corresponding points must be changed in order to improve this model and extract from it the three modes of variation.

The new algorithm was applied to a set of shapes and Figure 4.8 shows how the score is lowered as the algorithm progresses. The sharp change in values is the result of selecting different subsets, but the downward trend of the curve is clear to see.

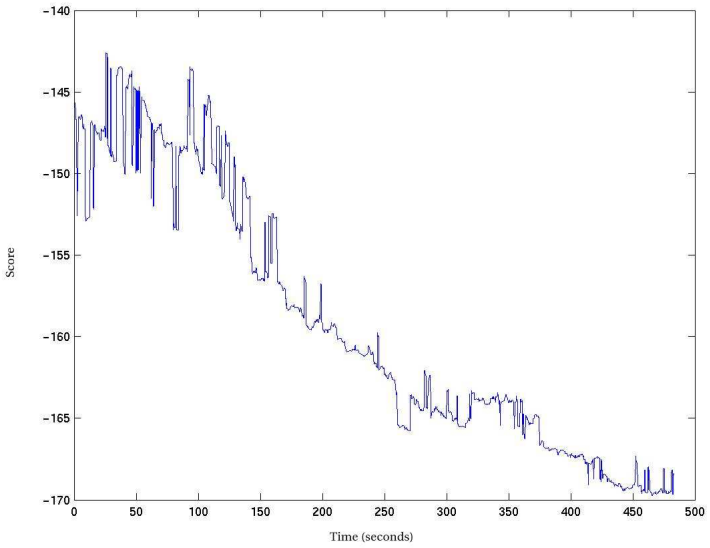

Figure 4.8: Shapes being aligned according to the description length, which is being minimised. A choice of a subset changes every 10 iterations. It is evident that the score goes lower.

4.6 Summary

This chapter has explained, but has not yet demonstrated, some of the advantages gained by using an MDL approach for choosing landmark points in a set of shapes. The notion of an objective function was explained, as well the idea of using an information-theoretic objective function. Once this function is in place, there are various issues that are concerned with the optimisation. The way by which good solutions are sought is rather crucial. Subsequent chapters, which cover detailed experiments, elaborate further, using more practical examples.

Chapter 5

Model-based Registration

"As order exponentially increases, time exponentially speeds up."

– *Ray Kurzweil.*

THIS chapter outlines an approach for constructing appearance models using the images alone, building on the work on automatic shape model building described in the previous chapter. A registration process is used to produce *dense* markup automatically, which delivers the needed correspondences for the process of model building. This exemplifies the reciprocity, as well as tight relationship, between the task of registration – that which establishes dense correspondence – and models that utilise this correspondence.

5.1 Overview

An important component of the proposed framework is the construction of an appearance model from a NRR *automatically*, i.e. without manual intervention. The chapter begins by providing an explanation of a simplified registration framework that brings one dimensional data into a state of alignment. This is followed by a more detailed explanation about the way alignment is used to construct models of appearance directly.

Throughout this section, an approach is described for automatic construction of appearance models using a criterion related to *complexity*. However, in principle, one could replace this with other criteria, as indicated in benchmarks that subsequently appear. Technical details that are described here frequently refer back to earlier and similar work [25, 26]. That work backed the idea of building *shape* models which are progressively refined by assessing their complexity. The current work is distinct and it is a cumulative step owing to the inclusion of intensity data in the model. Finally, a method for evaluating such models can be described in a greater level of depth.

5.2 Applying the MDL Approach to Images

This section briefly explores the data used in a set exploratory experiments and algorithms that are involved in building good models from a group of images.

5.2.1 Image Data

It was our choice to concentrate on images that resemble the test data shown in the previous chapter, so an algorithm was created which synthesises a variety of bumps.

Synthetic Bumps

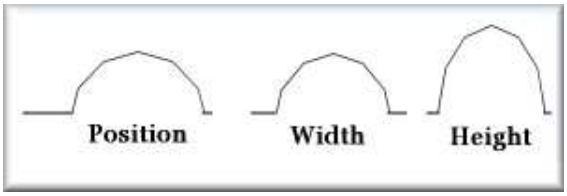

Figure 5.1: Examples of image data which illustrate of the three variation modes $\{p, w, h\}$. These variations characterise the data to be analysed. It is not the same image being resampled; rather, these are separate images.

The 1-D data which is dealt with hereafter is a simple elliptic bump. Several such bumps, once they have been generated artificially, vary in 3 distinct (yet loosely-related) ways, as shown in Figure 5.1. The data is

generated by moving along x (from the left side of the curve to the right) and considering 3 random variables, $\{p, w, h\}$, which lie within a sensible range that restricts variability. The variables correspond to:

- The point where the bump starts;
- The width of the bump (which defines where the curve flattens and the bump ends) and;
- The height of the bump.

These variables are referred to as p, w, and h, respectively.

Visualising 1-D Images

In order to show how 1-D images get registered, a set of representations is needed. The videos in the enclosed CD-ROM contain illustrations that are dynamic (time sequences) where warps are repeatedly applied, leading to change over time. Upon warping, the bump data can be perturbed by varying the positions of sample points (y-values), as illustrated in Figure 5.2.

Figure 5.3 shows another set of 6 different 1-D images visualised in 2-D (a cross-section). They are examples of what the data looks like in its most simplistic form, i.e. when only 2 points (the edges) are used to sample it.

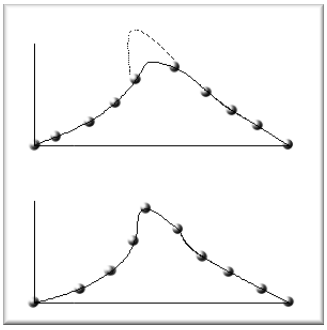

Figure 5.2: The top bump illustrates how an image, sampled using a set of points, can be transformed by movement of these points. This makes the bump at the top appear similar to the bump at the bottom.

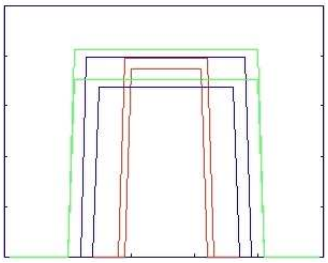

Figure 5.3: A simplified set of bump data. Difference between data instances is indicated by distinct colours (or shades). This simplified data was used in some of the earlier experiments.

Dealing with several such bumps[1] in turn, one can visualise them as shown in Figure 5.4 and Figure 5.5, where bumps are laid side by side. Many of the auxiliary videos depict a 1-D registration using this representation of the data. Figure 5.6 has these images displayed as surfaces.

[1]It should be emphasised that bumps are – at least in this context – merely 1-D images, or vectors.

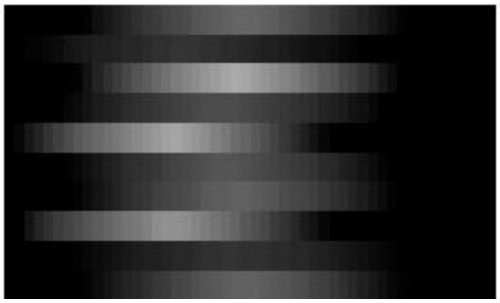

Figure 5.4: Bumps being registered. The registration process is visualised by an image composed of data vectors (laid horizontally alongside one another). The columns are 1-D vectors to be interpreted as grey-scale pixels.

Figure 5.5: A larger example of the pixel representation for 1-D bump data.

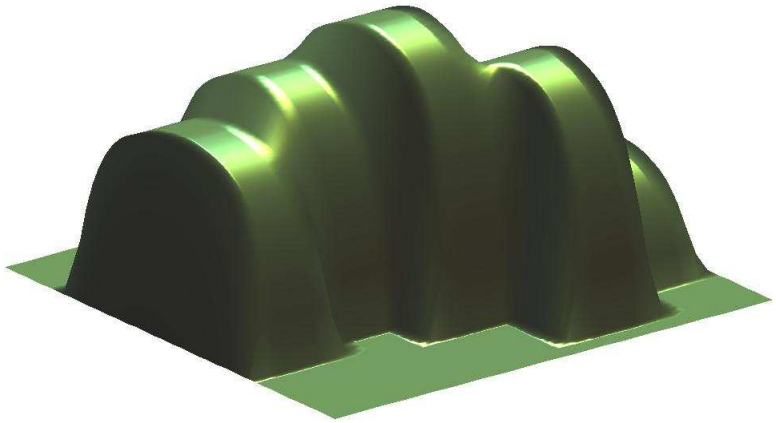

Figure 5.6: Original image set depicted in 3-D. A set of size 5 bump is shown before applying any warps which are intended to align this data.

5.2.2 Experimental Framework

With a theoretical approach to warping and a method for constructing test data, a framework was needed to test them and see how theory translates into reality.

A program with a graphical user interface was built in order to manage the task of registration and conduct experiments (see the screenshot in Figure 5.7). Thousands of these experiments have been automatically documented by the program, which became a popular public project. It is the second such program, the first one being an application that handles and manages optimisation of shape models (see Figure 4.5 on page 85).

The first step taken by the NRR program (called AART) is the synthesis (i.e. generation) of some bumps. It is a random process. These bumps varied in their height and width; the 'step size' of the bump was set to be fixed, i.e. the bump was initially flat at the top.

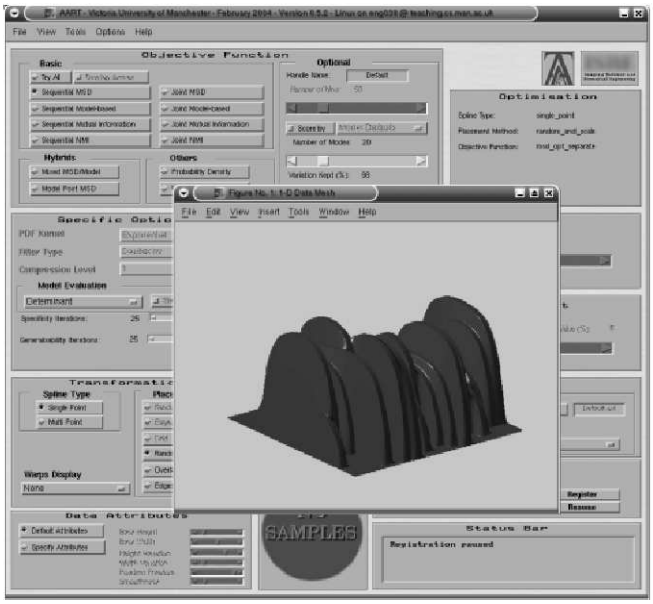

Figure 5.7: Autonomous Appearance-based Registration Testbed (AART): the program built to handle NRR and model building using just a GUI

5.2.3 Image Warping

Representation and Constraints

An original, unwarped image is function $f(\mathrm{x})$, where $f(\mathrm{x})$ is sampled at a set of equally-spaced points (which are the pixel set, x).

Suppose we have a warp $\phi(x)$ which collapses the unit interval [0,1] onto itself, with no folds or tears.

To transform images, we chose to employ clamped-plate splines as they address known flaws often encountered when thin-plate splines and the

B-splines are used. The clamped-plate splines prevent any of the regions in an image from being torn or folded, hence they preserve the existence and integrity of all image regions[2].

When the image is warped, using biharmonic clamped-plate splines, the set of previously equally-sampled pixel positions will move to new positions $\phi(x)$, and carry their image value with them. When this is then resampled onto equally-spaced pixels using linear interpolation, distortions can occur because an important part of the image can be missed totally when it is resampled. Two cases, namely single-point and multi-point warps, were both considered, where in the latter case composition is used to obtain more flexible deformations.

Manipulating Correspondences

Image correspondences can be identified by applying warps to the deformation fields that resample the images. There were a few challenges, however, but they are discussed later in this chapter.

Although the property of height was not intended to be ignored during registration (its signal is the maximum intensity, which cannot be ignored), it was expected that it would remain unchanged. This is due to the nature of CPS-based warps being used. They are perfectly diffeomor-

[2]Particularly in the bio-medical domain, visibility of all constituent structures becomes crucial.

phic[3][113], provided their constraints are obeyed. The warps themselves comprise a position and they are manipulated by altering the direction and magnitude of a single central knotpoint, or multiple knotpoints. Both cases were considered here and the warps were applied to localised parts of the line, not the entire line.

In the experiments described below, the bumps were all symmetric w.r.t. the center of the bump. The height, h, took a single value value from a fixed range set {*hi, low*} where $lo = 0$ and $0.7 < hi < 1$. The data was therefore far simpler than 1-D data which is not constrained in any way. The height of the bump and the position at which the bump goes high jointly defined the appearance that bump, so two real numbers (a tuple) at the minimum sufficed for the reconstruction of each bump.

5.2.4 Objective Function

The objective function used here uses an implicit similarity measure, which is an approximation of the quality of a model. To repeat a key argument, any set of unregistered images can be used to build such a model, but only a properly registered set of images – that which results in high groupwise similarity – builds a good model. This observation can be exploited to create a similarity measure that not only deals with *pairs* of images, but can also deal with large *sets*.

[3]While the bump may have its form tweaked and manipulated, its highest peak should be preserved although it may move leftward or rightward. This assumes that boundaries for the warp intensity are not affected.

Similarity is computed indirectly in this case. The algorithm does so by calculating the complexity of a combined model of shape and intensity, namely by looking at the covariance matrix of that model. To efficiently evaluate model complexity, we obtained $\sum_{i=1}^{n} log(\lambda_i + \delta)$ where $\lambda_{1<i<n}$ are the n eigenvalues of the covariance matrix whose magnitudes are the greatest. This approximates

$$det(\mathbf{M} + \delta) \equiv \prod_{i=1}^{n} (\lambda + \delta)_i \propto \sum_{i=1}^{n} log(\lambda_i + \delta) \equiv log(det(\mathbf{M} + \delta)) \tag{5.1}$$

where M is the model's covariance matrix under consideration and δ is a small constant which ensures we do not multiply by 0. The algorithm makes the registration purely model-driven so that no reference image is required. The objective function leads to one distinct solution without dependence upon individual images. This resolves the recurring issue of having to select a reference image and treat the problem as if it relies primarily on that one image.

5.2.5 Reparameterisation and Optimisation

Given an approximation of the model quality and a method for selecting warps, a general optimiser is used to find sets of warp parameters that maximize the quality of the model. In the next section, a few methods are described which can assist the optimisation by having it focused on interesting parts of the problem or by reducing the complexity if the problem,

in which case the optimisation can be made quicker.

5.2.6 The Registration Algorithm

This section presents the model-based objective function in the form of pseudo-code. Initially, the algorithm is demonstrated using simplistic one-dimensional data. The algorithm can be conceptually divided into two parts as follows:

Initialisation

- Generate images or retrieve them from a file.
- Optionally, apply image smoothing.
- Choose image reference. By default, the image closest[4] to the mean of all images gets selected.

Main Loop

- For a (pre-defined) number of iterations in the registration:
  - ✧ Set the level of precision for the optimiser to reach. Ideally, this level should be increased (or tolerance lowered) when advancements toward the goal made are small, i.e. registration is approached.

[4]Proximity is calculated based on the Euclidean distance, which identifies the reference that is, on average, nearest to all other images.

- ✧ Repeat for all images:
  - ❏ If the current image is not a reference image[5]:
    - ▷ Set up the positions of knot-points (random placements, drawn from a Gaussian distribution, are typically chosen).
    - ▷ Given the knot-points positions, apply warps to the current image and seek the warp parameters which minimise the complexity of the model built from the entire set of data.
  - ❏ end if
- ✧ end repeat

- end for
- Statistics and registration logging take place.

At the core of this algorithm lies an evaluator of the complexity of the model. The determinant, as in Equation 5.1, is used and shown in all experiments and results that follow. It aligns with the strategy used to build shape models, as described in Chapter 4. It is related to MDL, but it provides a good approximation of it. Simplicity is particularly important in this case because many models and built and evaluated while optimising.

[5]The reference must remain static in this case.

5.2.7 The Algorithm Visualised

Figure 5.8 shows the process which is outlined above. The framework is demonstrated in a simplified form in both cases (schematically and algorithmically). A reference image, as seen at the top of the figure, remains unaffected while all other images are manipulated. These are used to construct a model from which a the complexity measure can be derived, e.g. description length or just its approximation. Based on that measure of complexity, subsequent warps are applied to the group of images.

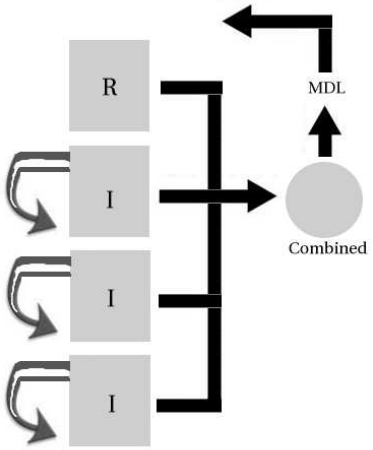

Figure 5.8: Schematic of the registration algorithm. A reference image (R) and the remainder of the warped set of images (I) form a combined model (circle) which is evaluated in an MDL-like fashion to refine subsequent warps.

What follows are some key example experiments which, together with the corresponding videos (see accompanying CD-ROM), demonstrate the process of registration visually, with an element of progression (dimen-

sion of time). What follows are also some perils and solutions to problems encountered on the path towards a workable NRR algorithm that builds appearance models.

5.3 Initial Experiments - Registering Images

This section presents several examples from key experiments that demonstrate the value of the methods just described. These experiments provide the basis for an automatic construction of appearance model, as described in the next section.

In Chapter 4, a set of experiments were shown where one automatically finds parameterisations that align shapes. Herein, images are treated in 1-D where intensity is being analysed in a similar fashion to shapes. Preliminary experiments deal with simple data and are used to learn a variety of lessons. Only the take-home messages, along with some visual results and quantitative graphs, are listed here.

5.3.1 Model-based Registration Experiments

A set of experiments were used to show that when the NRR objective function is model based, it can then achieve its goal quickly and effectively. Figures 5.9 and 5.10 show that when registering a set of 10 images by minimising the determinant of their model, the score is lowered

significantly at the start, but stabilises as convergence is approached. Figure 5.10 shows the improvements in terms of MSD whereas 5.9 uses the model's determinant to illustrate the effects of running the optimiser (registering a set of bumps).

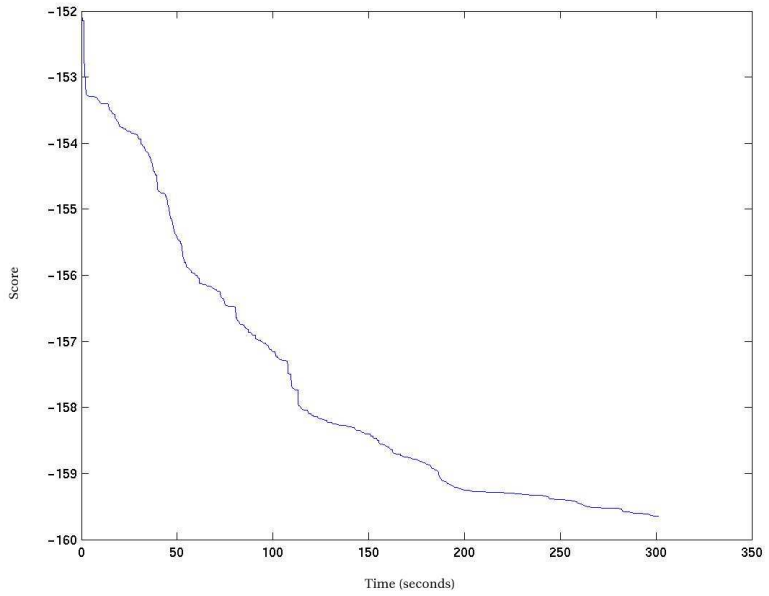

Figure 5.9: Images being registered according to an approximation of the description length of the entire set of 10 bumps. The X-axis corresponds to runtime time in seconds and the Y-axis is a logarithmically-scaled estimation of model complexity (determinant of the model).

A reasonable registration of a set of 10 small images can take a few minutes to complete on a modern PC. Figure 5.10, for instance, shows how a steep curve reaches a plateau after several thousands of warps were considered and applied to the images. Figure 5.9 gauges time and demonstrates the fact that convergence is gradually approached.

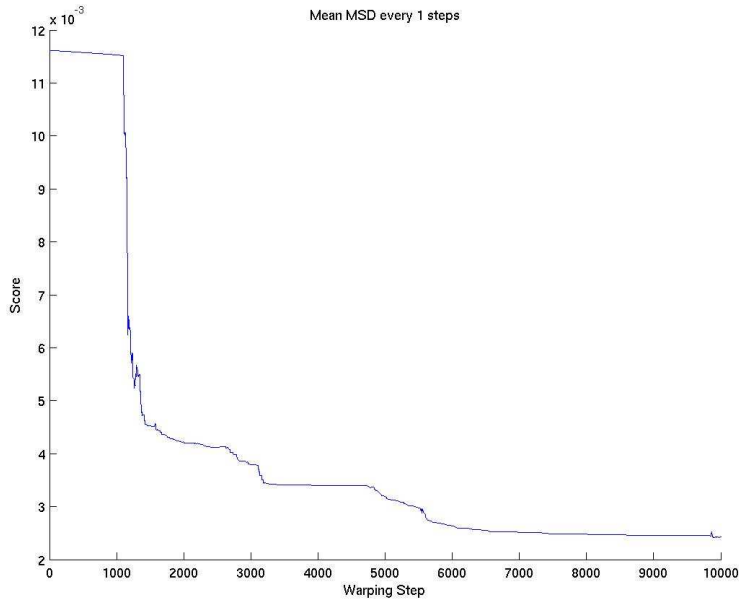

Figure 5.10: Mean MSD is measured at each point during the model-based registration of 10 bump images

Curiously enough, as Figure 5.9 demonstrates, there are some situations where very few improvements are made to the quality of the model (indicated by the flatness of the curve). This occurs due to the nature of the algorithm, which can get temporarily caught in local minima, in which case many warps that are considered simply get discarded. To improve the performance of the search for good warps, a better optimisation method was sought. For this reason, at one stage of the work, taboo search (TS) was considered.

Taboo search [36] facilitates a framework where there is danger of being drawn to incorrect solutions during optimisation. Taboo search is

a technique that attracted some interest in the 1990s. It makes use of knowledge about the search space while optimising. Thus, it can look up previous decisions and reach good solutions rather rapidly. It is similar to Simulated Annealing from a theoretic point-of-view.

5.3.2 Comparison of Registration Methods

Work on the model-based objective function was fruitful. Images of bumps were registered using a variety of registration methods, including the model-based method. Results are shown in Figure 5.11. Algorithms have their results displayed as curves and they are plotted as a function of iterations (number of warps). In order to adopt and use just one common measure that makes them comparable, registration quality is quantified in the figure using a measure of model complexity (the determinant). The methods include:

PDF - an NRR algorithm which considers the distribution (pdf) of intensities. It accumulated the scale of the distributions of intensity values located at each fixed image pixel in all the images. Consider, for instance, a situation where bumps are almost properly aligned, in which case the distribution of intensities across all images, at all positions along the X axis, should be small.

Wavelets - a method which is commonly used for compression purposes, such as JPEG, and can study the variation of image intensities. The

use of wavelet functions is related to PDF, but compressibility – rather than distribution – is used. Several different wavelets can be used and about a dozen were considered and tested. They are related to the MDL principle because they give insight into the cost of encoding the set of images as a message. Each corresponding position in the image is considered for all images and this position gets its pixels compressed in order to study the level of variation at that position.

MI and NMI - these are the two mutual information-based measures described in Chapter 2. They are used to compute the informational overlap between a reference image and each of the other images in the set.

MSD - mean of the squared differences is used in a similar fashion to the above methods. A reference image is chosen which is nearest to the mean of all images.

Mixed MSD - a combination of both the model-based approach and the MSD-based approach (hybrid). The two methods are alternated to explore the possibility of escaping local minima by considering two separate objective functions.

Post-MSD - the model-based approach, followed by the MSD-based approach. MSD is used in the second half of the registration process, however complicated, long or extensive it may be.

The existing algorithm was well behaved. The problem as whole, however, needed to be thoroughly understood because warps were permitted to 'erode' parts of the bump image, thereby allowing bumps to grow shorter or nearly vanish altogether. Specifically, given the single- or multi-point clamped-plate splines, once a warp is applied to the image which is resampled, the limits in resampling can have the top of the bump eliminated. Since results are carried on for another warp to be considered, this resampling problem becomes greater at each iteration. The next subsection discusses this further and provides an example.

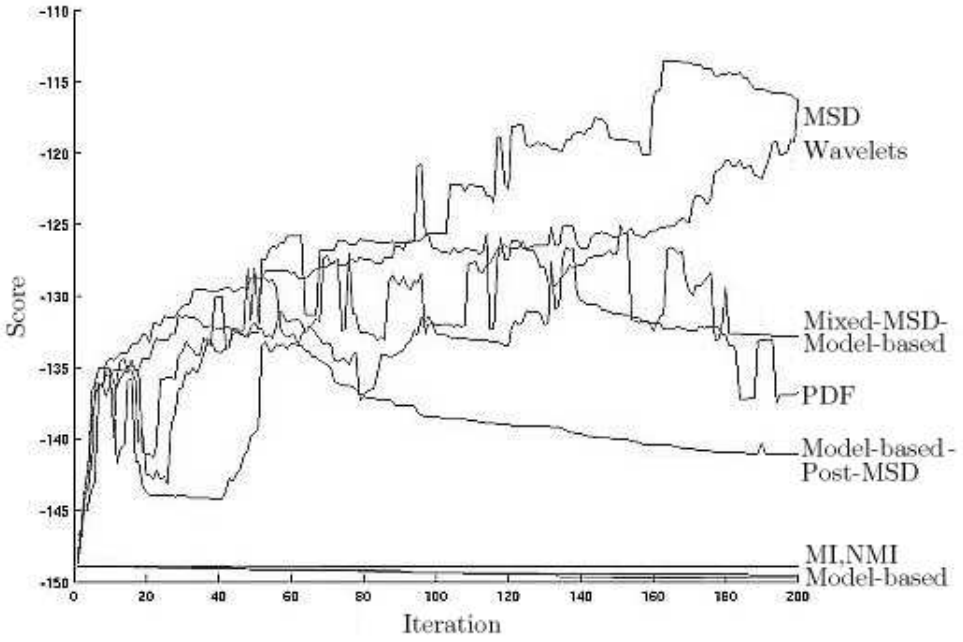

Figure 5.11: A comparative analysis of different objective functions. It illustrates the fact that the model complexity decreases only for the newly-proposed objective functions. The Y axis value is an indicator of model complexity, based on its determinant.

5.3.3 Deficiencies of MDL-based Models

It was later realised that model residuals must be included in some form or another (e.g. description length) in the objective function. As one example of the need for this issue to be resolved, see Figure 5.12. When not accounted for properly, the quality of the model can be perceived as though it surpasses the point of optimality. This was actually known when applying the same type of approach to shape, so things were set up so that the residuals were always smaller than some specified amount that was irrelevant. The incomplete term used for a proper description length is described in [25].

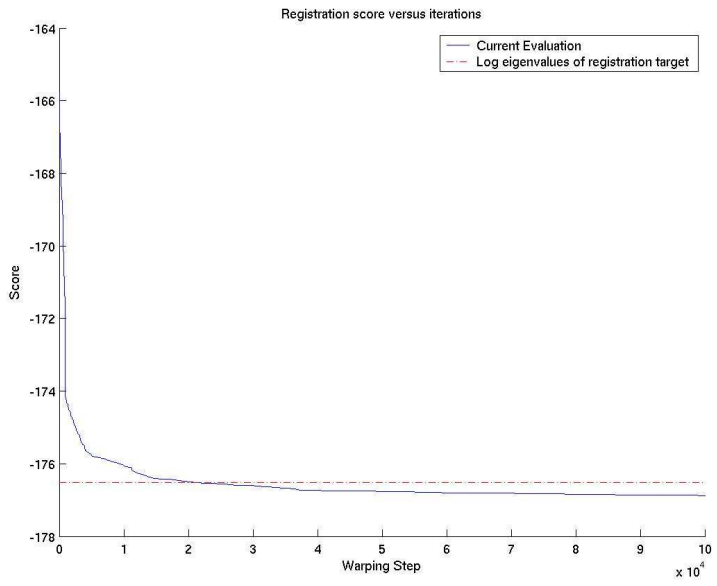

Figure 5.12: A long optimisation with the successful model-based algorithm shows that it surpasses what is questionably the correct solution (indicated by the red dotted line).

5.3.4 Comparison of Optimisation Regimes

Different optimisation regimes were investigated, e.g. by changing optimiser parameters for the standard optimisation function, as well as investigating approaches that are altogether different. An example of this is presented in the next section on automatic model building in 2-D.

5.4 Experiments

A group of experiments explored the behaviour of the algorithm and determined ways in which parameters or additional methodologies affect its performance. The experiments were all managed by the graphical front end, AART, which was deployed on 10 PCs at the clusters of the Computer Science Department. Hardware specifications used was kept identical so as to get identical results where time is being measured.

NRR comparisons were performed which yielded comprehensive numerical comparisons of NRR with different methods for assessing similarity across the sets of 1-D image. The optimisation, for example, was altered in terms of its granularity and subsets were used to build a model on which the similarity measure is based.

I will show an NRR based on the whole set racing against a subset approach, as described in Chapter 4, a comparison which involves changing the sensitivity of the optimiser, and another comparsion which looks at

the weighting assigned to shape and intensity when building combined models of shape and intensity.

To give an overview, there are three series of experiments where things vary systematically in order to learn their effect on the overall NRR performance. I consider:

1. Weighting (shape/intensity)
2. Optimiser tolerance
3. Subsets

Models built using different approaches were difficult to discern visually, unlike the quality of the NRR. So, having chosen to compute the correct solution, I was able to assign a figure of merit (the determinant in this case), which was based on the ground truth.

5.4.1 Weighting

As a start, I performed 10 repeated experiments with 10 different random instantiations of 10 bump images (comprising 200 sample points on each). These were registered using a model-based objective function and 200 iterations were used to refine the model.

The goal was to demonstrate that by changing the tradeoff between shape and intensity (relative scaling) in the construction of the model, not only

are different models obtained at the end, but the NRR too is improved or degraded. In order to show that one weighting scheme trumps another, I consider the correct solution – that which brings perfect alignment. This solution is studied based on the synthetic data at the start and then compared to at the end. In addition, examples of the variation were given by showing separately the shape variance, intensity variance, and value of the objective function which is based on the combined models.

20 modes of variation were considered when computing the determinant and 98% of the variation kept when applying PCA.

The experiments make use as their input 10 non-identical starting points with 10 different datasets that are imported from files. I performed the experiments with 10 different datasets and the at the very end of the registration measured:

- Model score
- Mean of shape variance
- Mean of intensity variance
- MSD score
- Time taken to run the experiment

Model score is the value of the objective function at the end of the NRR. It was also useful to measure is the variance computed for each mode of

variation (wI consider 20 modes in total). As NRR progresses, I find that intensity variance is decreased (bumps are better aligned), whereas it is typically increased for shape because we have different warps applied to each bump.

These variances are good reflectants of the modes of the resultant model, broken down into shape and intensity (a combined one depends on weighting).

One could measure something else, but the models are good representations of the shape and intensity components which directly relate to image dissimilarity. At the moment, from the final shape model it is also possible to retrieve (reconstruct) each shape and then measure, using MSD, how it compares to the original shape. This might not be valuable to these experiments, but the funcationality is incorporated into the code and the GUI.

Figure 5.13 and Figure **??** show the results from this first batch of 10 experiments which study shape and intensity weighting. On the X axis, the scaling is shows where 100% means only intensity is accounted for, whereas 0% means that only shape variation is accounted for.

It is possible to see that, in terms of the objective function, a low value is reached when shape and intensity are considered in isolation. This can be explained by the variation of shape and intensity. When only shape is accounted for (0%), then it's optimised to the point where its final variance is low, whereas it is high for the intensity component. Conversely, when

intensity is optimised, then the images reach good alignment, whereas the shape (wrap) component becomes more irregular.

5.4.2 Optimiser Tolerance Section

In this set of experiments, the aim is to compare result obtained not by considering different optimisers (although this is paractically possible), but to look how the refinement with single-point CPS warps is handled when varyious degress of tolerance are considered.

This study involved running a set of experiments where the NRR is applied to a fixed set of bumps (as above) and the results then assessed at the end, in the very same way as before. Then, while varying the tolerance and rerunning the experiment it's possible see the effect on the quality of the results. I repeat this experiment 10 times with a different set of bumps and each time then consider the average to demonstate that one choice of tolerance is more favourable than others.

The numbers used are as above (200 iterations, i.e. 200 passes through the set). To tolerance, changes are made by scaling by an order of magnitude each time, 3-10 decimal places, plus 12 and 15.

Figure 5.15 and Figure 5.16 show the results from these experiments. With this general-purpose Nelder-Mead optimiser, the best results with respect to the ground-truth solution are obtained when a 6 decimal point accuracy is required from the optimsier. As expected, the greater the

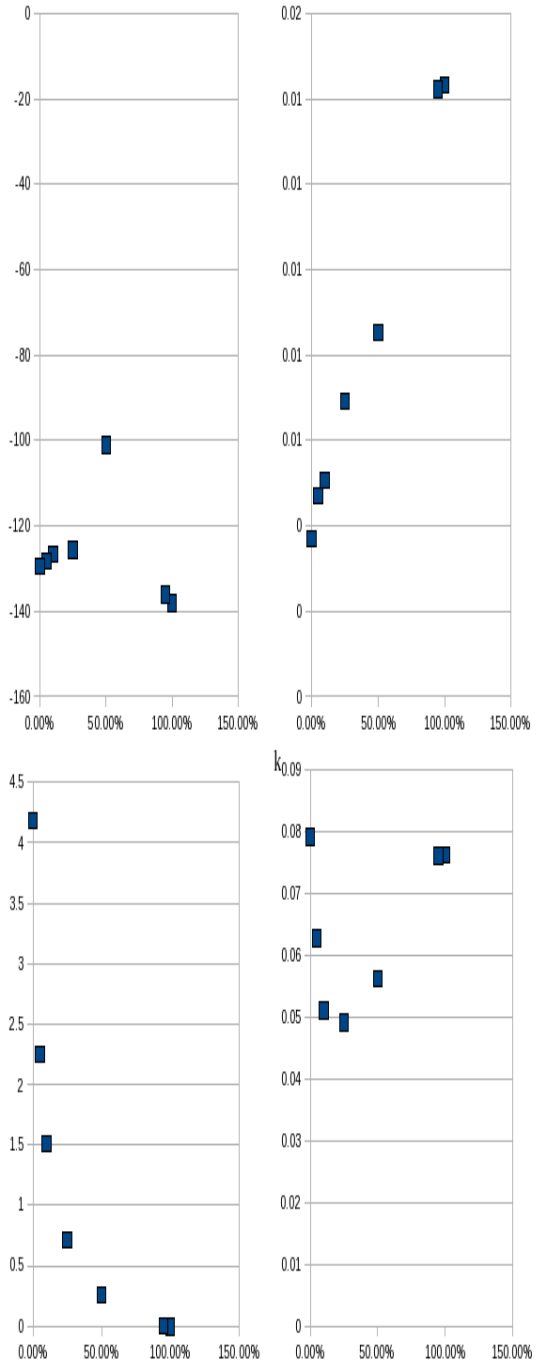

Figure 5.13: Top-left: final model score as a function of shape/intensity weighting (lower is better); top-right: MSD score; bottom-left: mean of shape variance; bottom-right: distance from the correct solution.

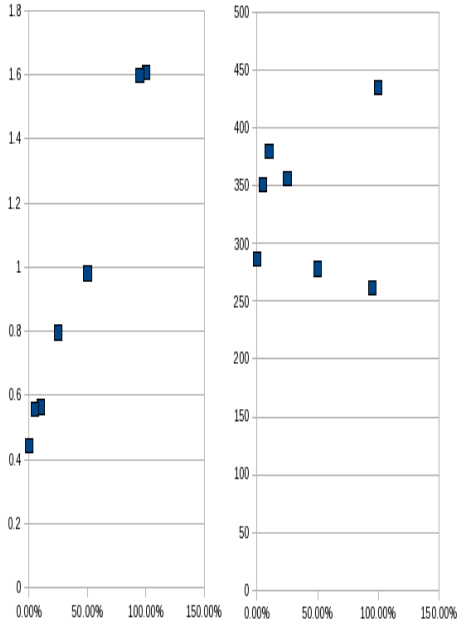

Figure 5.14: On the left: mean of intensity variance as a function of shape/intensity weighting (lower is better); on the right: mean time (in seconds) to complete one NRR

accuracy required, the longer it takes for the NRR to be completed. Requiring too great an accuracy at each stage has the negative effect of not allowing the algorithm to explore distant (and potentially better) solutions, so choosing too high a tolerance has can lead to deviation from the correct solution.

5.4.3 Subsets

Given a set of parameters that seemed to work well in the previous experiment, I performed the following last set of experiments.

10 repeated experiments with 10 different random instantiations of 10 bumps (comprising 200 sample points each) were performed. It needed to be demonstrated that by considering a subset of all images, the determinant becomes small or the same value reached more quickly (given an objective function which is model-based). 20 modes of variation were considered when computing the determinant and 98% of the variation kept when applying PCA.

200 iterations should be used to refine the model and subsets re-chosen in random (without repetition in selection) every 10 iterations.

Figure 5.17 and Figure 5.18 show the results from these experiments. It's possible to conclude that while the approach has helped very little in improving the results, which are only marginally different.

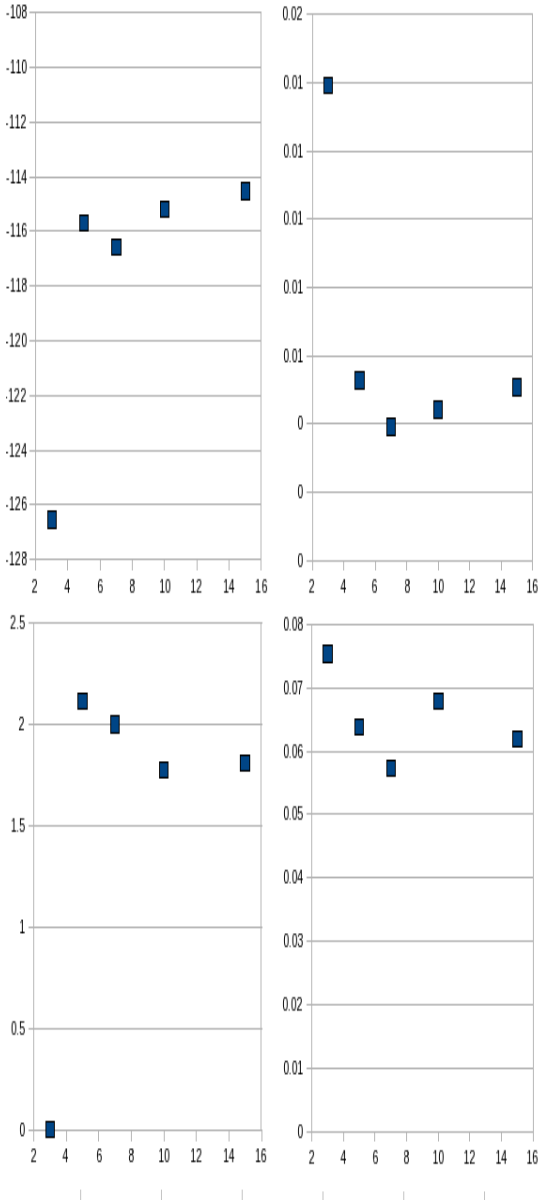

Figure 5.15: Top-left: final model score as a function of optimiser tolerance (lower is better); top-right: MSD score; bottom-left: mean of shape variance; bottom-right: distance from the correct solution.

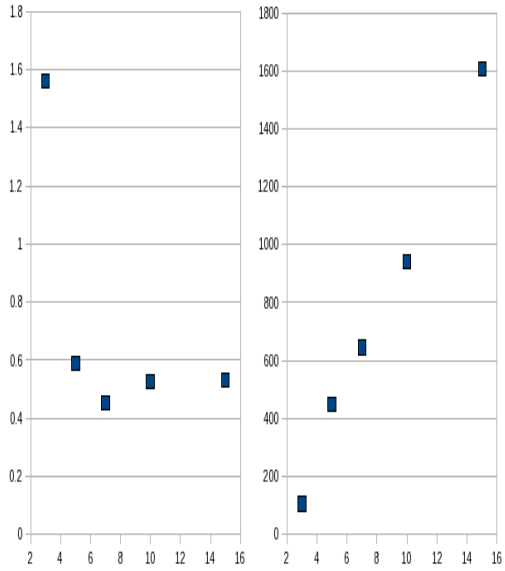

Figure 5.16: On the left: mean of intensity variance as a function of optimiser tolerance (lower is better); on the right: mean time (in seconds) to complete one NRR

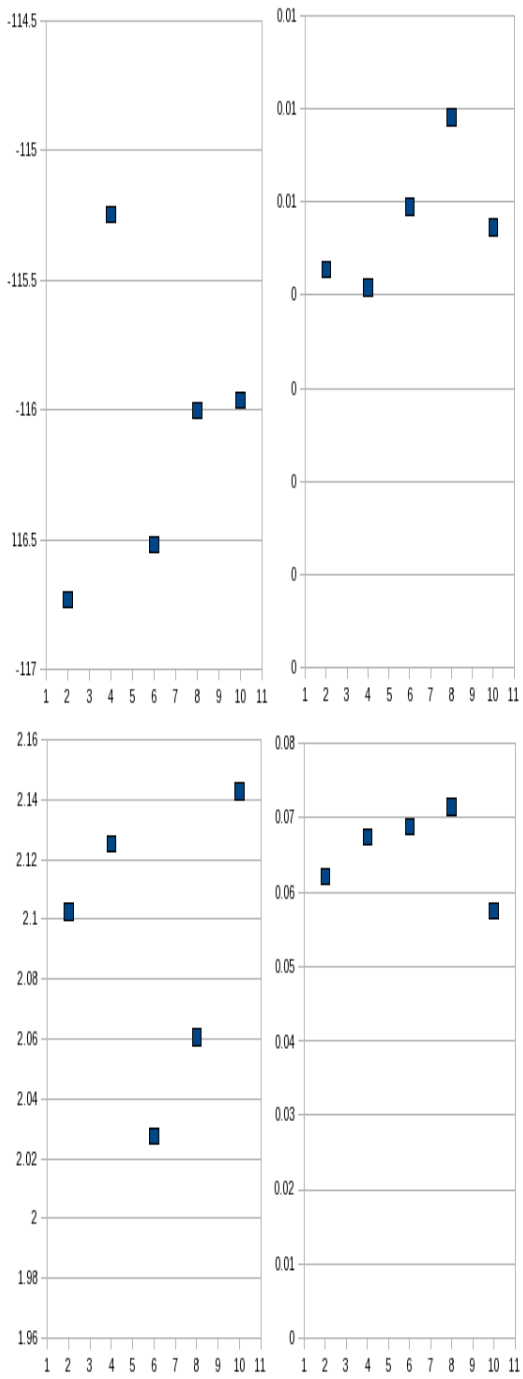

Figure 5.17: Top-left: final model score as a function of the size of the subset (lower is better); top-right: MSD score; bottom-left: mean of shape variance; bottom-right: distance from the correct solution.

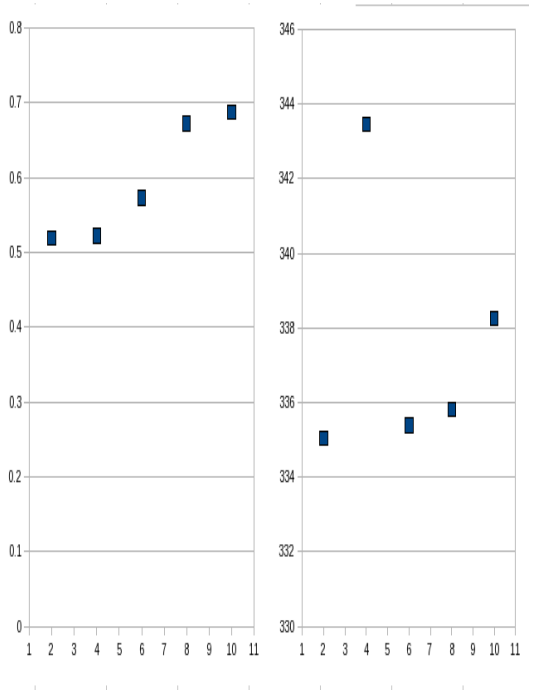

Figure 5.18: On the left: mean of intensity variance as a function of the size of the subset (lower is better); on the right: mean time (in seconds) to complete one NRR

Other experiments that may be interesting to learn from are once which study how varying the number of model modes considered or variation kept when applying PCA affects overall performance. In a problem such as this, there are many different parameters that can be varied simultaneously, so fine-tuning of the algorithm is an interesting task. In practice, I found that changing the assignment of value to particular parameters depending in the progress of the algorithm was a way of producing pleasing results.

5.5 From Registration to Models

The previous sections covered a variety of methods for registering 1-D data and a few challenges that are posed. The output of these registrations are warp fields which define one-to-one correspondences of points embedded in the data. Having obtained these correspondences, models can be built directly as described in § 3 on page 53. Because the correspondence is dense, there is no limit in principle on the number of sample points from which the appearance model is constructed.

Figure 5.19: The first (top) and second (bottom) modes of variation in the appearance model constructed from the 1-D bump data. The first mode shows the whole bump moving from side to side and the second mode shows it moving and spreading.

5.5.1 Automatically Building Appearance Models in 1-D

The work described in the remainder of this chaper was done by myself, the author, as well as colleagues at the department. It later led to further exploration in separate direction.

What a registration algorithm establishes, regardless of the objective function used, is dense correspondence in the images. Given this correspondence – no matter how it was obtained – it is possible to learn the variation in terms of shape and intensity. By applying the same algorithm to a set of one- or two-dimensional data, an appearance model can be built. The videos labelled `8.avi` and `14.avi` (see the accompanying CD-ROM) show the result of 1-D registration of bump data using a model-based objective function. A static image which corresponds to what the videos demonstrate is shown in Figure 5.19. Combined models are depicted, along with shape and intensity models.

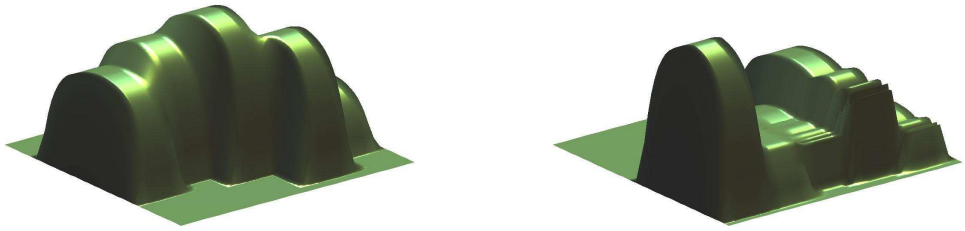

Figure 5.20: On the left: Example bump data is displayed in its initial form. Each bump on the surface represents a 1-D vector. On the right: Example of the early results of registration (200 iterations). The edges of the bumps clearly begin to align.

To demonstrate the advantages gained by the model-based approach, we also experimented with data where the correct solution is known. Generated data again depicted bumps and the sets of images were stochastically generated with significant variability that makes the problem challenging. We define a solution to be good when we observe proper alignment of the bumps and a resulting registered set that is distinct from any of the original images. At the same time, we are continuously delivered statistical models (as shown in Figure 5.20) of variable bumps. A combined model is derived from the shape model and the intensity model. Well-founded ways exist to visualise and evaluate them and it can be seen that the combined model is refined in the process, even after as little as 200 iterations.

After only a few minutes, good alignment amongst all bumps was ob-

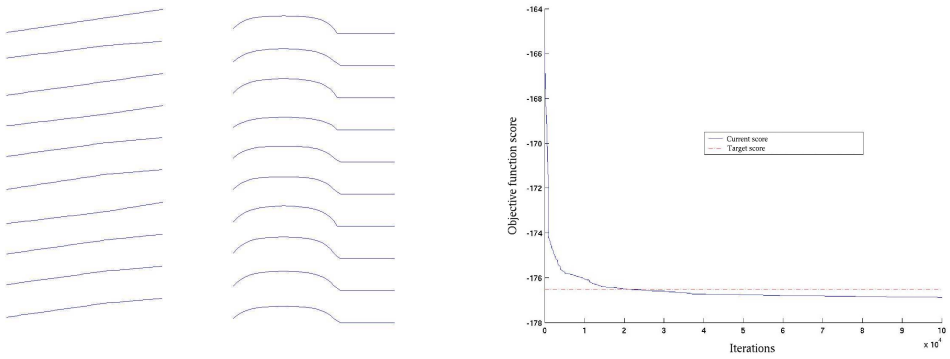

Figure 5.21: On the left: The correct warps that align given data and that same data with the warps applied. On the right: The value of the objective function as convergence is approached.

tained (see Figure 5.21). Sets comprising dozens of bumps could be successfully handled by the algorithm and a statistical model of their appearance emerged as a by-product of registration. When compared with the known correct solution, the quality of registration was high. It also successfully surpassed naïve implementations of some conventional algorithms.

As well as a basic model-based objective function, in this experiment we also investigated the use of subsets, as described in Section 4.5, to speed up the process. Subsets are chosen stochastically every 100 iterations, therefore the problem is simplified and the algorithm becomes more effective in dealing with large sets. It is worth adding that choice of warps was random at all stages, so no data-bias or *a-priori* knowledge was involved.

See Figure 5.22 on the next page and, once again, compare this with the

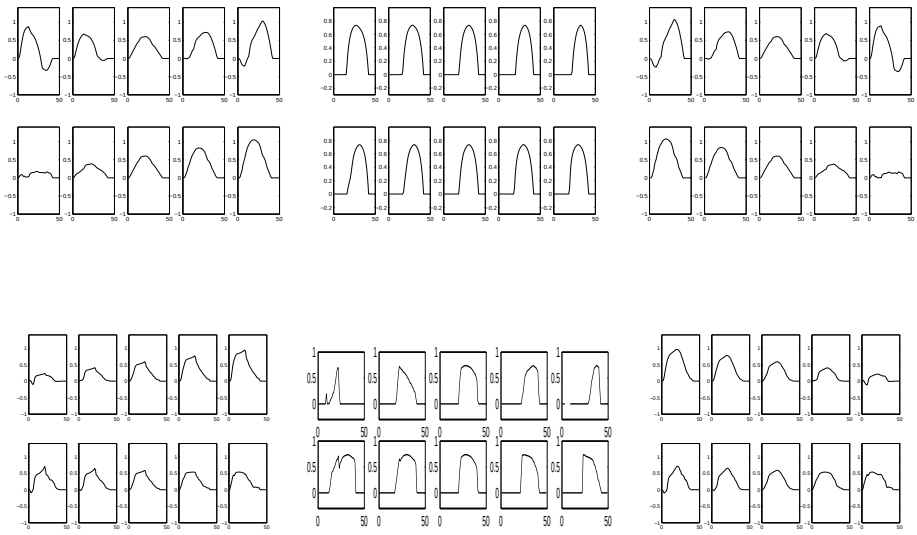

Figure 5.22: At the top, from left-to-right: Combined, shape, and intensity models of 10 data instances at the start. The principal modes are shown with up to ± 2 standard deviations away from the mean. At the bottom: corresponding models after registration.

data shown in Figure 5.20. The results suggest that our approach works properly while addressing common difficulties. It can handle large sets and provide a solution that does not depend on any arbitrary selection of images.

5.5.2 2-D Model Construction

Taking a similar approach and applying it to 2-D image sets, the same type of models can be built from brain data (see, for example, Figure 3.4 on page 69).

This short section explores the extension of the image registration method

to 2-D. Later comes a derivation of models from the registered data. It is important to emphasise that this one particular goal was achieved by a group of people, so it should not be counted as an achievement made solely by the author [22]. The author of this thesis primarily contributed towards making NRR work better, but not automating model building in 2-D, as derived from NRR.

The Objective Function

In order to obtain a set of corresponding points, transformation of data is needed. Two separate 'families' of registration algorithms were repeatedly used. The first method uses sum-of-squared-differences to measure similarity, whereas the second uses the minimum description length (MDL) framework. In both cases, transformation was handled solely using bi-harmonic clamped-plate splines. Although mutual information is a commonly-used similarity measure, it did not prove to be as fruitful during the experiments described in the remainder of this thesis.

Figures 5.24 and 5.23 demonstrate the registration process, essentially by overlaying images to get a mixture that is a chequerboard. It is a simple case of visually evaluating the differences between two images.

As part of the experiments which investigate optimisation through minimisation of the objective function, a survey was conducted to compare 3 methods. Figure 5.25 visually (as opposed to quantitatively) compares

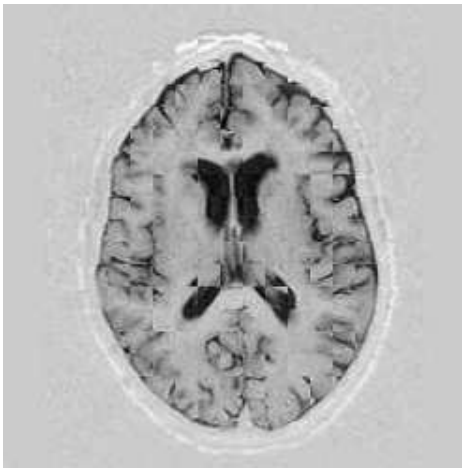

Figure 5.23: Chequerboard image showing the difference between two registered images. The objective function used was sum-of-squared differences.

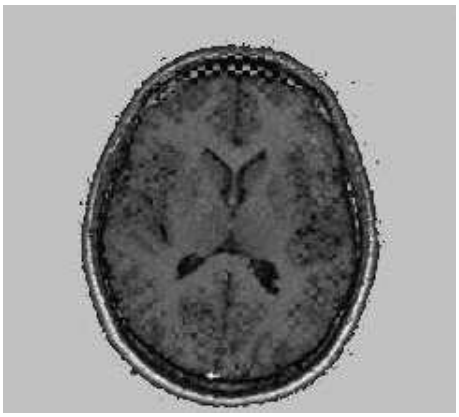

Figure 5.24: Chequerboard image showing the difference between two registered images (different from the image set shown in the previous figure). The objective function used here was mutual information.

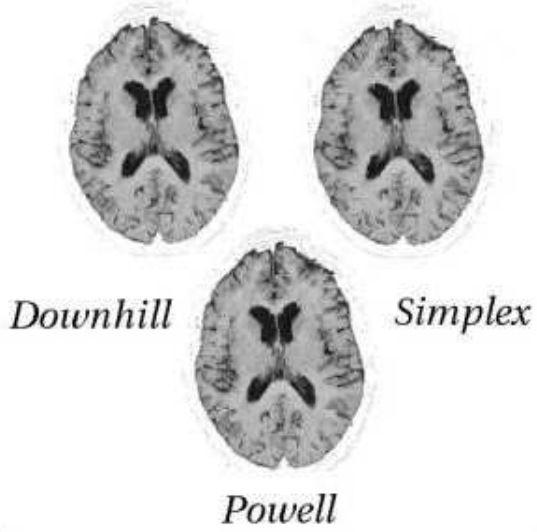

Figure 5.25: A survey of different registration optimisation methods. The figure shows the results in the form of chequerboard images, each for the results of a different optimisation method.

the level of refinement reached using the different optimisation methods studied. While the figures show similar results, the time taken to arrive at these result varied significantly. Powell was particularly slow compared to a simple downhill optimisation, which was used in the majority of cases at different scales[6].

The final product of this joint effort was a framework that transforms a registration, encoded as a set of deformation fields, into a model. The main challenge to address was finding transitions between the representation of warps (including triangulation) into a point-to-point correspondence that can be interpreted and consumed by the existing model

[6]Coarser images were registered first in order to get a good initial approximation for the warps and more sophisticated methods were used towards the end, when finer-level images were registered.

construction framework. Example result, which is an automatically-built model, is shown in Figure 5.26.

This key development not only proved that models can be built without manual markup, but it also showed that the resultant models are reasonable. It is important to stress that, due to the scale of this problem, no model-based objective function was used in the 2-D case. In principle, however, as demonstrated in 1-D, it should be possible use any reasonable objective function. There is no reason, other than scalability, why the method adopted for model construction ought not to involve model compactness as its criterion.

The Model Construction Algorithm

A registration algorithm is used to establish a set of deformations. The algorithm's complexity and its type are arbitrary. A method can then be used to build a default generative model of the set of deformations. The space of this model can be explored and searched. This approach, however, suffers from a couple of drawbacks:

1. The use of a default model will assume that Gaussian distributions are the ones which best fit the pattern of the data. This means that in the process of building the model, there would be a tendency to choose deformation which lead to Gaussian distributions.

Figure 5.26: The first mode (± 2.5 standard deviations) of an appearance model built automatically by group-wise registration.

2. A computational limitation is introduced because when deformation fields, from which models get built, are changed, their effect on these entire model leads to a cyclic situation, as described in [119] and also in Chapter 4 on shapes.

5.6 Summary and Discussion

This chapter demonstrated a general approach and framework for building appearance models directly from NRR. A registration algorithm can be used to identify correspondences in a group of images. In turn, these correspondences are used to build models of shape and appearance. While there are challenges to address, such as the speed of the optimisation and accuracy of the results, the method proves to be satisfactory enough to produce a model which correctly obtains modes of variation that were introduced artificially. At the core of this algorithm, there exists an approach to estimating model complexity – the determinant of an appearance model. This method is fast, but it does not have a principled theory.

Subsequent experiments were carried out which followed the path towards automated model construction. They built upon this work on 1- and 2-D and were used by colleagues who extended the ideas further and produced publications, e.g. [22]. The author of this thesis meanwhile proceeded to the problem of assessing NRR, which is the topic that the next few chapters cover in detail.

Chapter 6

Assessment of Non-Rigid Registration

"Nothing in life is to be feared. It is only to be understood."

– *Marie Currie.*

THE most valuable contribution of this thesis is the introduction of a framework wherein NRR can be assessed without the use of ground truth. The framework is valuable not only because NRR applications are routinely used but because it offers a solution to a problem where cumbersome annotation is otherwise needed [16].

The new assessment method exploits the relationship between NRR and model building, transforming the problem of assessing the NRR of a set

of images into that of assessing the quality of a model built from the registered images.

6.1 Overview

The key idea to rely on is the fact that a good NRR produces correspondences that build a good model. If good NRR algorithms and results lead to a good model, then all that remains is to evaluate models.

Previous work on shape models utilised measures of model specificity and generalisation. Chapter 4, which is focused on shapes, illustrates the fact that quality of shape models can be evaluated directly from just a set of points, but it is not the case when handling models of both shape and intensity because their significance needs to be weighed and then combined somehow. The contribution here is the introduction of Generalisation and Specificity, which are the means for evaluating appearance models. Both depend on generating a distribution of images using the model, then comparing the distribution with the distribution of training images. In turn, these two measures become metrics that tell apart good NRR from a poorer one.

6.2 Measuring Model Quality

A good model of a set of training data should possess several properties. Firstly, the model should be able to extrapolate and interpolate effectively from the training data, to produce a range of images from the same general class as those seen in the training set. We refer to this property as *generalisation ability*. Conversely, the model should not produce images which cannot be considered as valid examples of the class of image modelled. That is, a model built from brain images should only generate images which could be considered as valid images of the brain. We call this the *specificity* of the model. In previous work, quantitative measures of *specificity* and *generalisation* were used to evaluate shape models [26]. The extension of these ideas to images (as opposed to shapes) is presented here. Figure 6.2 provides an overview of the approach. It shows how clouds of points can be used to derive Specificity and Generalisation (as also shown in Figure 6.1 and explained later in this chapter) in a Monte-Carlo fashion. Twining and Taylor recently showed that, by using KL divergence, one can extract the same type of measures based on a sound theoretical basis [117].

6.2.1 Specificity

Consider first the training data for the model, that is, the set of images to which NRR is applied. Without loss of generality, each training image can

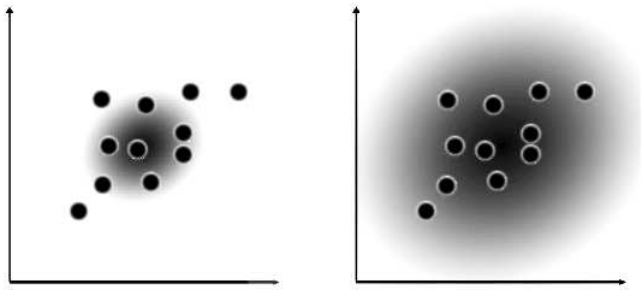

Figure 6.1: Training set (points) and model pdf (shading) in image space. **Left:** A model which is specific, but not general. **Right:** A model which is general, but not specific.

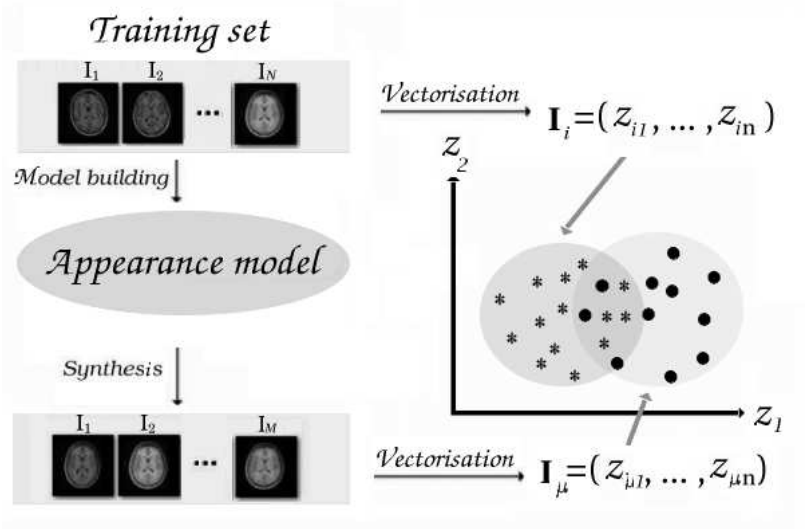

Figure 6.2: The model evaluation framework: A model is constructed from the training set and used to generate synthetic images. The training set and the set generated by the model can be viewed as clouds of points in image space ($\mathbf{I}_i$ represented by stars, and $\mathbf{I}_\mu$ represented by dots).

be treated as a single point in an n-dimensional image space. A statistical model is then a probability density function (pdf) $p(\mathbf{z})$ defined on this space.

To be specific, let $\{\mathbf{I}_i : i = 1, \ldots \mathcal{N}\}$ denote the $\mathcal{N}$ images of the training set when considered as points in image space. Let $p(\mathbf{z})$ be the probability density function of the model.

A quantitative measure of the *specificity* S of the model is defined, with respect to the training set $\mathcal{I} = \{\mathbf{I}_i\}$ as follows:

$$S_\lambda(\mathcal{I}; p) \doteq \int p(\mathbf{z}) \mathbf{min}_i \left(|\mathbf{z} - \mathbf{I}_i|\right)^\lambda \, d\mathbf{z}, \tag{6.1}$$

where $|\cdot|$ is a distance on image space (see Section 6.3), raised to some positive power λ (for the remainder of this chapter only the case λ = 1 will be considered). That is, for each point z on image space, the nearest-neighbour to this point in the training set is found. Then, the nearest-neighbour distances raised to the power λ are weighted by the pdf $p(\mathbf{z})$. Greater specificity is indicated by *smaller* values of S, and vice versa. In Figure 6.1, diagrammatic examples of models with differing specificity are given.

The integral in equation 6.1 can be approximated using a Monte-Carlo method. A large random set of images $\{\mathbf{I}_\mu : \mu = 1, \ldots \mathcal{M}\}$ is generated, having the same distribution as the model pdf $p(\mathbf{z})$. The estimate of the

specificity (6.1) is:

$$S_\lambda(\mathcal{I};p) \approx \frac{1}{\mathcal{M}} \sum_{\mu=1}^{\mathcal{M}} \mathbf{min}_i \left(|\mathbf{I}_i - \mathbf{I}_\mu|\right)^\lambda, \tag{6.2}$$

with standard error:

$$\sigma_S = \frac{SD_\mu \left\{ \mathbf{min}_i \{ |\mathbf{I}_i - \mathbf{I}_\mu|^\lambda \} \right\}}{\sqrt{\mathcal{M}-1}}, \tag{6.3}$$

where SD_μ is the standard deviation of the set of μ measurements. Note that this definition of S does not require that the space of images is constructed. Instead, one simply needs to be able to define distances between images. This is discussed in Section 6.3 below.

It is worth adding that while more specific models will be close to the training data, complete separation between the two groups of images (training and synthetic) ensures there is no bias.

6.2.2 Generalisation

A measure of generalisation is defined similarly, simply reversing the direction of the nearest-neighbour distance measure:

$$G_\lambda(\mathcal{I};p) \doteq \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \mathbf{min}_\mu \left(|\mathbf{I}_i - \mathbf{I}_\mu|\right)^\lambda, \tag{6.4}$$

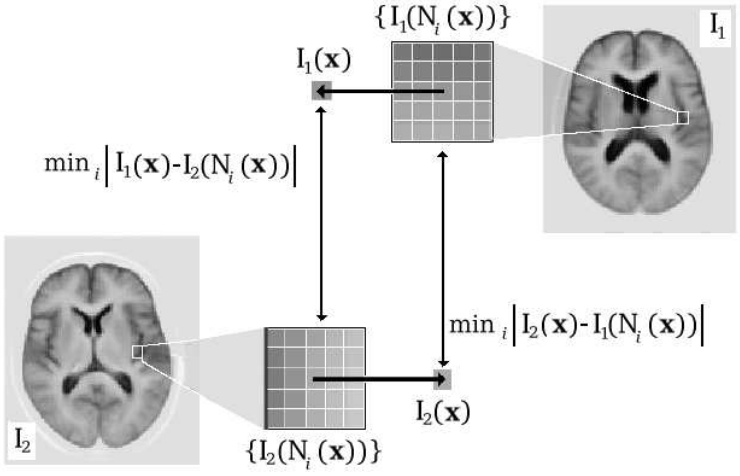

Figure 6.3: The calculation of a shuffle difference image

with standard error:

$$\sigma_G = \frac{SD_i\left\{\mathbf{min}_\mu\{|\mathbf{I}_i - \mathbf{I}_\mu|^\lambda\}\right\}}{\sqrt{\mathcal{N}-1}}. \tag{6.5}$$

That is, for each member of the training set $\mathbf{I}_i$, the distance to the nearest-neighbour in the sample set $\{\mathbf{I}_\mu\}$ is computed.

Note here that both measures can be further extended, by considering the sum of distances to k-nearest-neighbours, rather than just to the single nearest-neighbour. However, the choice of k would require careful consideration and in what follows, experiments shown are restricted to the

single nearest-neighbour case.

6.2.3 Caveat

An assumption is made when defining specificity and generalisation in this way. The pdf is required to be of a form which permits meaningful distances to be measured and one must consider the fact that it is important to show the distinction between a measure which spans a large space and one which is centered only around the training set. For example, generalisation might not properly reflect on the distribution among training images.

The situation is likely to be acceptable if the sets being considered are quite large. It is not possible to generate or obtain more training images (as opposed to synthetic ones), but assuming that the training set is sufficiently large, as it is in subsequent chapters, the difference between the sizes of the two sets – training and synthetic, that is – can blur this gap to a degree. It is thus recommend that this method is applied where a distribution of the training set is fairly consistent, without many outliers. Large training sets are recommended also.

Another important point to be made is that while the measured specificity and generalisation increase as a function of misregistration, this does not necessarily mean that they measure misregistration. In a similar vein, methods that are used in an NRR objective function can improve a registration, but it must not be assumed that they are absolute measures of

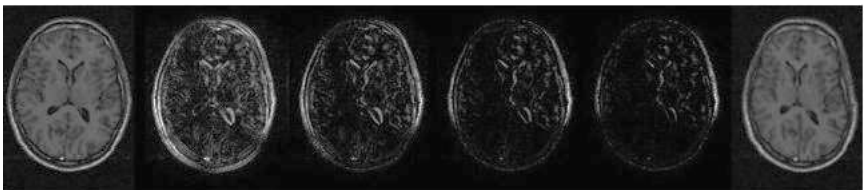

Figure 6.4: A comparison between shuffle difference images evaluated using various size neighbourhoods (radius r). **Left:** original image, **right:** warped image, **centre, from the left:** shuffle difference images with $r = 1$(Euclidean), 1.5, 2.9 and 3.7 pixels.

image similarity or indicative of registration quality. For example, mutual information-based NRR may lead to a good registration, but more evidence would be needed to show that mutual information can also be used to compare the quality of several different registrations. The same word of warning applies here, but as I will show later, an empirical relationship is demonstrated between assessment which is based on ground truth and the method presented in this chapter.

6.3 Image distance measures

The definitions provided for specificity and generalisation (equations 6.2 and 6.4) require a measure, $|\cdot|$, of separation in image space. The most straightforward way to measure the distance between images is to treat each image as a vector formed by concatenating the pixel/voxel intensity values, then take the Euclidean distance. This means that each pixel/voxel in one image is compared against its spatially corresponding

pixel/voxel in another image. Although this has the merit of simplicity, it does not provide a very well-behaved distance measure since it increases rapidly for quite small image misalignments. Another possibility is to use Image Euclidean Distance (IMED), as proposed in a recent paper [123], by considering the two images as surfaces $(\mathbf{x}, \mathbf{I}(\mathbf{x}))$ and searching from each position $\mathbf{x}$ in one image along the tangent to the surface at point for the nearest point on the surface covered by the other image. While this approach was implemented and tested, it was not considered further due to efficiency and time constraints.

This observation led to consideration of an alternative distance measure, based on the 'shuffle difference', inspired by the 'shuffle transform' [55]. Given two images $\mathbf{I}_1(\mathbf{x})$ and $\mathbf{I}_2(\mathbf{x})$, the shuffle distance between them is defined as

$$D_s(\mathbf{I}_1, \mathbf{I}_2) = \frac{1}{n} \sum_{\mathbf{x}} \mathbf{min}_i \|\mathbf{I}_1(\mathbf{x}) - \mathbf{I}_2(\mathbf{N}_i(\mathbf{x}))\| \tag{6.6}$$

where $\|\cdot\|$ is the absolute difference, there are n pixels (or voxels) indexed by $\mathbf{x}$, and $\{\mathbf{N}_i(\mathbf{x})\}$ is the set of pixels in a neighbourhood of radius r around $\mathbf{x}$.

The idea is illustrated in Figure 6.3. Instead of taking the sum-of-squared-differences between corresponding pixels, the minimum absolute difference between each pixel in one image and the values in a neighbourhood around the corresponding pixel is used. This is less sensitive to small misalignments, and provides a better-behaved distance measure. The tolerance for misalignment is dependent on the size of the neighbourhood

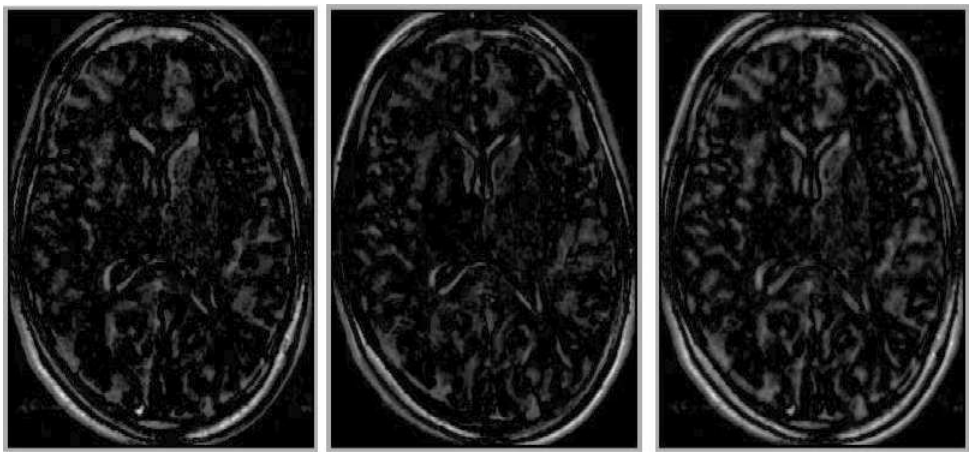

Figure 6.5: Examples of the shuffle difference image: from first to second (left), from second to first (centre), and the symmetrical shuffle difference image (right)

(r), as is illustrated in Figure 6.4.

It should be noted that the shuffle distance, as defined above, depends on the direction in which it is measured (see Figure 6.5), hence is not a true distance. It is trivial to construct a symmetric shuffle distance, by averaging the distance calculated in both directions between a pair of images. It was found, however, that the improvement obtained in this way was not significant, and did not justify the increased computation time. In what follows, the asymmetric shuffle distance is used exclusively.

6.4 Summary and Discussion

This chapter showed how the problem of assessing NRR could be transformed into one of measuring model quality. It then introduced an approach which enables distances between images to be measured. It de-

fined Specificity and Generalisation as measured of appearance model quality based on a measure of inter-image distance. It introduced a practical method of measuring inter-image distance and showed that, given distance measures and variety of images derived from a model, the quality of this model can be evaluated.

The measures used to evaluate models are Specificity and Generalisation. By considering images that are non-rigidly registered and building a model from them, it is then possible to assess NRR, using Specificity and Generalisation.

The next chapter utilises this new framework and explores its behaviour. Systematic experiments are used to show that the method works. Although the approach outlined above is principled, it still requires experimental validation and this can be achieved by measuring the change in model quality as the correspondence of an initially-registered set of images is progressively perturbed. The results are then compared with those obtained using a state-of-the-art assessment method, based on the overlap of ground-truth pixel/voxel labels. The results demonstrate that, not only is the proposed approach capable of assessing NRR reliably without ground truth, but that it also provides a more sensitive measure of misregistration than the overlap-based approach.

In Chapter 8, the new method is also applied to compare the performance of different registration algorithms on several sets of MR images of the brain, demonstrating that the method is able to discriminate between dif-

ferent methods of registration in a practical setting. The methods developed for measuring model quality are also valuable – in their own right – for other purposes. They can be used to compare different methods of model building [95].

Chapter 7

Validation Methodology and Experiments

"Nothing is so simple that it cannot be misunderstood."

– Jr. Teague.

In this chapter, NRR of a brain dataset is studied in order to show that Specificity and Generalisation are valid measures of the degree of misregistration of a group of images. It is expected that, as registration is degraded, Specificity and Generalisation should change accordingly. Experiments were performed to validate the approach for evaluating NRR, making use of ground-truth annotation to provide an independent measure of misregistration.

Taking a set of registered images, for which ground-truth labels were

available, a series of controlled deformations were applied. This produced progressively-increasing misregistration and made it possible to investigate how measures of Specificity and Generalisation varied, as a function of the known misregistration. Generalised overlap (see 2.3.2) was also measured for each of the deformed image sets, using the ground-truth labels, to provide a comparison.

7.1 Method

7.1.1 Image Data

To conduct the experiments, a set of MR images of the brain was used. The set, which will be referred to as the 'MGH Dataset' (see Acknowledgements), was a set of 2-D transaxial mid-brain slices, extracted at an equivalent level from each of a set of affinely aligned T1-weighted 3-D MR scans of $\mathcal{N} = 36$ normal subjects. As well as the images themselves, there was access to ground-truth data, in the form of dense (pixel by pixel) anatomical label maps for the grey and white matter, the caudate nucleus, and the lateral ventricles. These labels were further divided into left and right hemispheres. The anatomical labels were obtained by manual annotation under conditions of rigorous quality control (details available at www.cma.mgh.harvard.edu). An example image and the corresponding label maps are shown in Figure 7.1.

The set of images was non-rigidly registered using a Minimum Description Length (MDL) NRR algorithm [119], and this registration was used as the starting point for a systematic evaluation of the effects of misregistration. Although the initial registration cannot be assumed to be perfect it is reasonable to assume that each deformation degrades the registration rather than improving it.

7.1.2 Perturbing the Initial Registration

In order to perform a systematic evaluation of the effects of misregistration, multiple image sets were created, based on the MGH Dataset, but with controlled degrees of misregistration. To create a misregistered set, the original image set was taken and had applied to it a set of smooth pseudo-random spatial warps, based on biharmonic Clamped Plate Splines [118]. The warp for each image was controlled by 25 randomly placed knot-points, each displaced in a random direction by a distance drawn from the positive half of a zero mean Gaussian with SD σ, which controls the degree of misregistration. This provided a very general family of warps. The degree of misregistration was quantified by measuring d, the magnitude of pixel displacement averaged over the whole image. A total of 70 misregistered image sets were generated – 10 warp-set instantiations for each of 7 different values of d (0.0643, 0.249, 0.685, 1.36, 2.21, 2.76, and 4.15 pixels). Examples of warped images are shown in Figure 7.2.

The sets of chosen warps were diffeomorphic by definition, so they preserved the topology of the images while the magnitude of localised warps ensured that there was no problematic distortion. The spatial spread of these small warps was assured to occupy different regions of the image and thus affect most image pixels as well. No warps were applied next to image boundaries, so as to avoid pixels from being scattered and pushed outside the image.

I tested a variety of sets and observed the effect of the warps on images in order to ensure that warped images did not lose their natural topology. The intensity of the warps applied was quantified by values of 10, 40, 110, 220, 370, 460, and 670. Rather than provide arbitrary numbers which are algorithm specific, I refer to pixels-quantified units in the experiments that follow.

The scales of deformation were carefully chosen such that they provided more insight into the behaviour of the experiments at points near the correct solutions. The deformations at the lower end of the scale were greater in number and sample points in the graphs likewise. The greatest level of deformation was not exceeded in these experiments because greater deformation did not seem to result in brain images whose appearance looked quite so reasonable. A deformation too great would explore situations that are less likely to be encountered in practical applications.

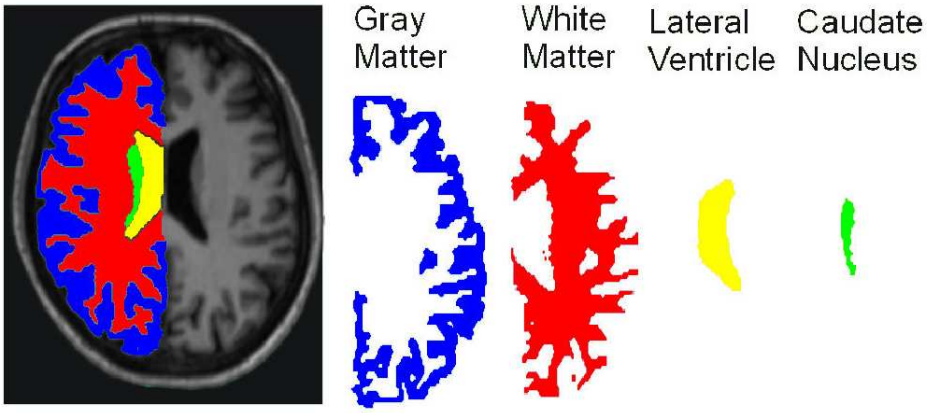

Figure 7.1: An example affinely-aligned brain image and its accompanying anatomical labels, both overlaid and expanded, for grey matter, white matter, the lateral ventricles, and the caudate nucleus. The labels are also divided into left and right.

7.1.3 Validation using Warped Images

Given the 70 image sets described above, each with known average misregistration, d, the relationship between d and Specificity, Generalisation, and Generalised Overlap was investigated, by calculating the mean and standard error for each measure over the 10 warp instances for each value of d. In total, there were 71 image sets to study. One is the original registered set and the other 70 image sets comprise 10 instantiations for each value of d.

It is worth emphasising that no images are actually being registered here. Instead, sets are generated by simulating different levels of misregistration and then studying this misregistration. Each set of 10 instantiations is derived from the same image, but the process is stochastic, so the instantiations are unique.

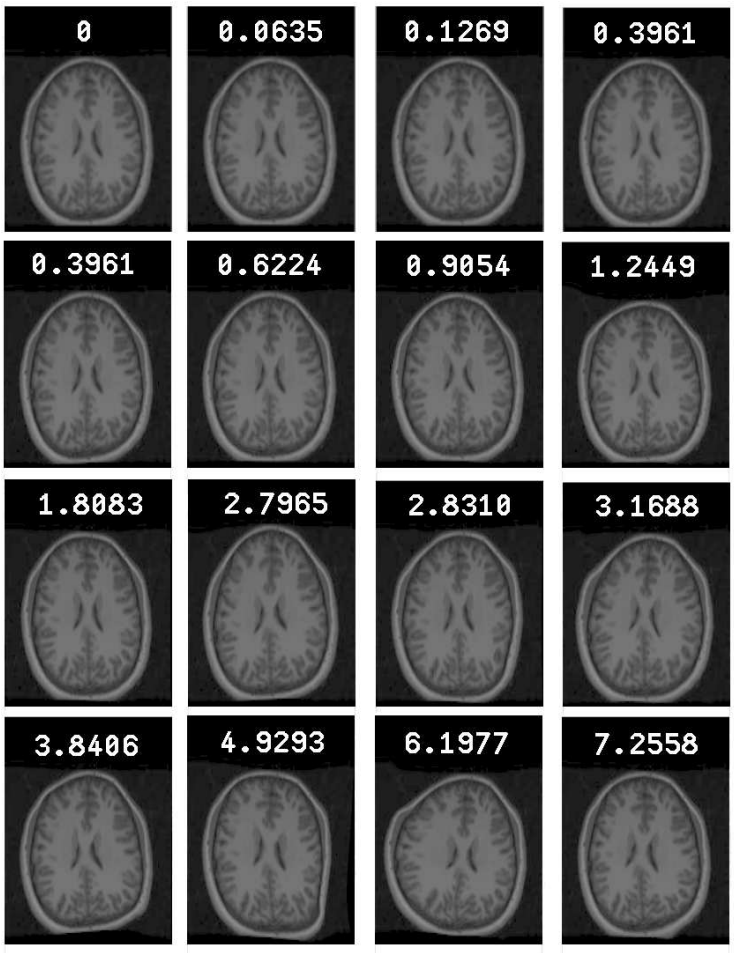

Figure 7.2: An original image from the MGH Dataset (top left) and examples of warped versions of the same image obtained using different values of d, the mean pixel displacement (shown on each image).

For each misregistered image set, Specificity and Generalisation were calculated, as described in Section 6.2, using m = 15 modes of variation for the model and $\mathcal{M}$ = 1000 synthetic images drawn from a Gaussian model distribution. This was repeated for values of shuffle radius, r, of 1 (Euclidean distance), 1.5, 2.1 and 3.7, as defined in Section 6.3, corresponding to circular neighbourhoods contained within 1x1, 3x3, 5x5 and 7x7 pixel patches respectively. These experiments were repeated with 2.5%, 5.0% and 10% Gaussian intensity noise added to the misregistered images, in order to investigate the sensitivity of the model-based measures to image noise. This makes it possible to argue in defence of the robustness of these measures to noise. Figure 7.3 shows an example brain image with varying degrees of noise added.

With the ground-truth annotation, Generalised Overlap with volume, equal, inverse volume and complexity weightings were calculated, as defined in equation 2.1. Different values of α_l affected the different components of that equation. The mean and standard error for each measure over the 10 warp instances for each value of d was also calculated.

7.1.4 Sensitivity

The size of perturbation that can be detected in the validation experiments will depend both on the change in the values of the measures as a function of misregistration and the standard error of those values. To quantify this, the sensitivity of a measure was defined as follows.

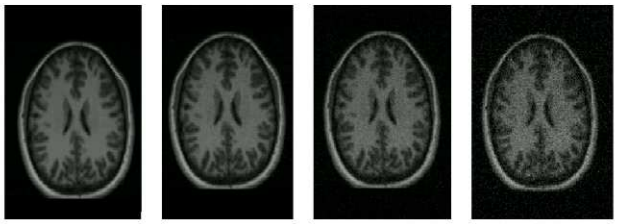

Figure 7.3: The effect of varying the level of noise applied to the original image (left). Noise levels of 0%, 2.5%, 5%, and 10% (right) are shown.

$$D(m;d) = \frac{1}{\sigma_m}\left(\frac{m(d) - m(0)}{d}\right), \tag{7.1}$$

where $m(d)$ is the value of the measure for some degree of deformation d, σ_m is the standard error of the estimate of $m(d)$. $D(m;d) = 1$ is the change in d required for $m(d)$ to change by one noise standard error, which indicates the lower limit of change in misregistration d which can be detected by the measure. D is a function of d; to simplify comparison between different methods of evaluation, we also use the mean sensitivity over a range of values of d.

In order to compare the sensitivities of different methods of evaluation, the expected error in D also needed to be estimated. Since the validation experiments provided repeated estimates of $m(d)$, one can obtain empirical estimates of the errors in $m(d)$, $m(0)$, and σ_m. These can be combined, using error propagation in Equation 7.1, to estimate the uncertainty in the estimate of sensitivity.

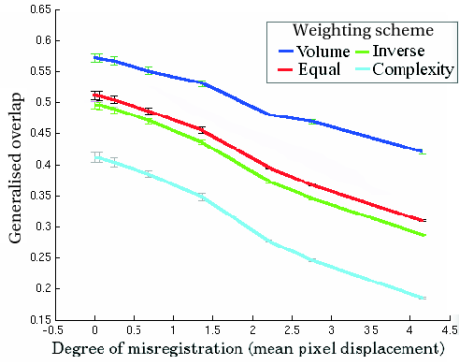

Figure 7.4: Overlap measures (with corresponding $\pm$ one standard error errorbars) for the MGH dataset as a function of the degree of degradation of registration correspondence, d. The various graphs correspond to the various tissue weightings as defined in Chapter 2.

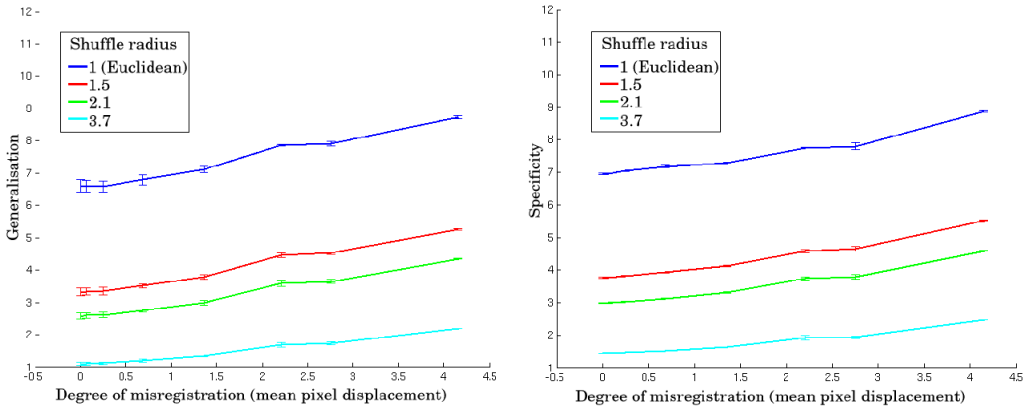

Figure 7.5: Generalisation & Specificity for various definitions of image distance (varying shuffle radius) with corresponding $\pm$ one standard error errorbars as a function of the degree of degradation of the registration correspondence d for the MGH dataset

7.2 Results

Figure 7.4 plots each of the four variants of the generalised overlap measure, as a function of d, the degree of misregistration. As expected, the value decreases monotonically with increasing misregistration in each case. This shows that the two gold-standard measures of misregistration (mean pixel displacement and ground-truth overlap) are in agreement, which validates the experimental framework. The measure of overlap is also proportional to the level of distortion that simulates and corresponds to misregistration, which validates the perturbation method.

The measure of overlap is also proportional to the level of distortion that simulates and correponds to misregistration, which validates perturbation method.

Similarly, Figure 7.5 plots Generalisation and Specificity as functions of d, for different values of shuffle radius r. The results are qualitatively similar to those obtained for generalised overlap, except that both measures *increase* monotonically with increasing misregistration, as expected (see Section 6.2). These results show that, over the range of misregistrations investigated, the model-based measures are good surrogates for d, the mean pixel misregistration. Since the warps used to introduce controlled misregistration were of very general form, there is no reason to suppose that this result is dependent on the pattern and specifics of misregistration.

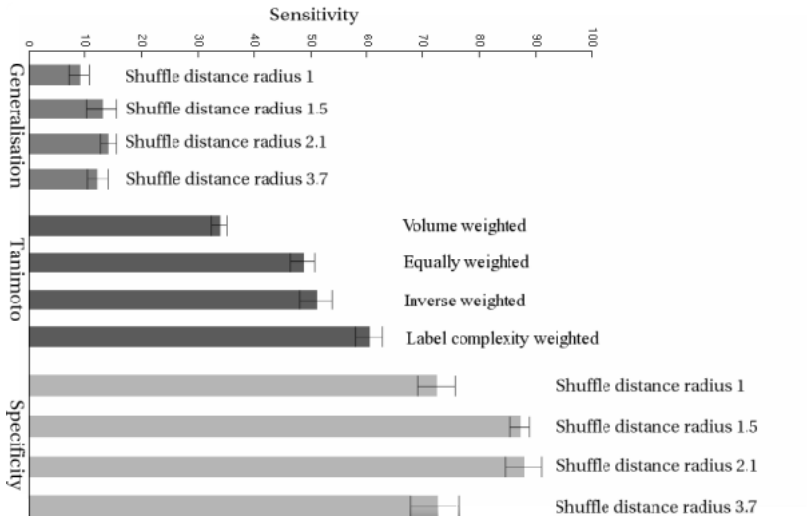

Figure 7.6: Mean sensitivity of different NRR assessment methods over the full range of deformations d, shown with $\pm$one standard error errorbars

7.2.1 Sensitivity

Figure 7.6 shows the results of applying sensitivity analysis to the validation study. These demonstrate that Specificity is more sensitive (i.e. is able to detect smaller misregistrations) than the overlap-based approach, which is in turn more sensitive than Generalisation. Note from the error bars that these differences are statistically significant. Maximum sensitivity is achieved with a shuffle radius of 1.5 or 2.1. Figure 7.6 also suggests that the most sensitive generalised overlap measure is obtained using label-complexity weighting. The choice of the Tanimoto Coefficient is consistent through all the experiments and each fuzzy overlap measure is based on it. The Tanimoto Coefficient is monotonically related to the popular Dice Similarity Coefficient (DSC), which would therefore produce

similar results.

7.2.2 Effect of Noise

The validation experiments were repeated and sensitivity analysis reported above with added image noise. Although the absolute values of the model-based measures were shifted upwards, as would be expected, there were no changes in the relative values, nor any systematic or statistically significant changes in sensitivity, even for 10% added noise. Figure 7.7 shows that as the degree of noise increased, Specificity is still able to detect significant differences in terms of the level of misregistration.

7.3 Discussion

A correlation was shown between the degree of misregistration and measures such as Specificity and Generalisation. Specificity was found to be more sensitive than Generalisation, which is unsurprising given the difference between the size of the training set and the size of the set derived from the model. As pointed out in the previous chapter, these two measures are separate but not independent. Both Specificity and Generalisation are needed in order to perform a full and valid assessment, but ways of combining these two to obtain a single figure of merit is not trivial. It is not possible to assess the quality of a registration using specificity

alone or using generalisation alone because in very special cases, there is a chance of the method failing slightly. For a good model to be built, those two measures would need to be optimised at the same time.

Despite this separation into two measures, in should be possible to consider just Specificity, but then use Generalisation to ensure an extreme case is not leading to an inaccurate conclusion. In all cases I have come across, there was a duality between those two measures, but Specificity was the more valuable among the two.

Overall, the experiments demonstrated, using a series of experiments, that the method is sensitive to reasonable degrees of misregistration and is robust to noise. It remains to be seen how the method can be applied to real-world problems, which is the topic the next chapter covers.

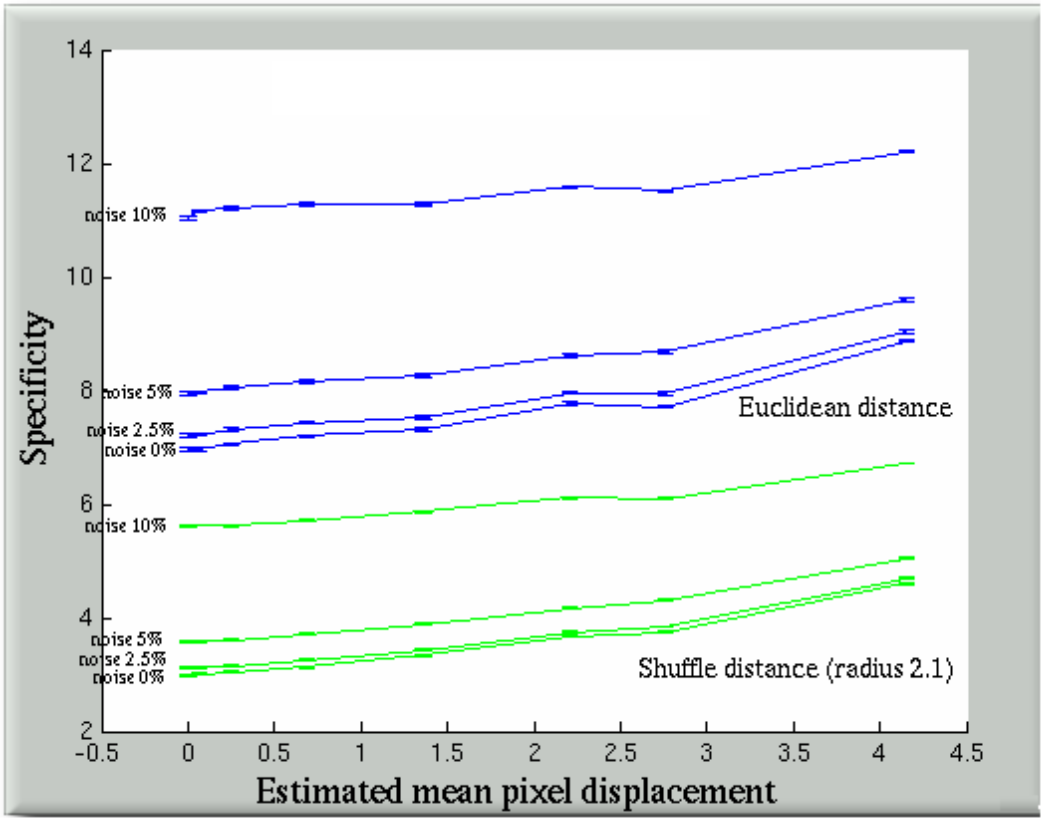

Figure 7.7: Specificity as a function of misregistration is shown for a number of datasets created by adding noise to a set the brain images (0%). The steepness of the curve is affected by high levels of noise, but the behaviour of the curve remains consistent.

Chapter 8

Application to Evaluation of NRR

"Civilisations can only be understood by those who are civilised."

– Alfred North Whitehead.

THIS chapter presents a comparison of different methods of NRR. A brain dataset is used to gauge the performance of one method which uses anatomical labels to measure overlap and shows that a model-based approach to assessment – one which does not require anatomical labels – can actually perform better.

The results from two separate experiments demonstrate behavioural consistency with different datasets. This shows a practical application of

what was described in the previous chapters and further validates the approach.

8.1 Comparing Registration Algorithms

To illustrate the application of model-based evaluation in practice, NRR results obtained using three different methods for registering a group of images were compared, as described below. The intent was to establish whether it was possible, in a practical setting, to detect significant differences in performance between different NRR algorithms. All three registration methods used the same piecewise affine representation of image warps [22] and the same multi-resolution optimisation framework. The same number of iterations (function evaluations) were used in each case.

The three registration algorithms were applied to two datasets. The first datasets, which is referred to as “MGH", was described in the previous chapter. The second set of images, which will be referred to as the ‘Dementia Dataset’, consisted of a set of 2-D transaxial mid-brain slices, extracted at an equivalent level from each of a set of affinely-aligned T1-weighted 3-D MR scans of $\mathcal{N} = 104$ subjects entered into a clinical study of dementia.

The MGH Dataset was used because it allowed the evaluation results obtained using Specificity and Generalisation to be compared with an evaluation based on the Generalised Overlap measure (using ground truth).

For these experiments $\mathcal{M}$ = 500 synthetic images were used to estimate Specificity and Generalisation. The Dementia Dataset was used because it was more representative of a typical clinical study, and it was important to demonstrate that the results were not dataset-dependent. For these experiments $\mathcal{M}$ = 1000 synthetic images were used.

The three registration methods used were as follows.

8.1.1 Pairwise Registration to a Reference

A commonly used approach to registering a group of images is to register each image in the group in turn to a reference image chosen from the group, using a pairwise objective function (e.g., [91]). This approach was used as a baseline, with a sum of absolute intensity differences objective function (which gave slightly better results than sum of squared differences or mutual information).

Pairwise approaches to registration can produce reasonable correspondences, but suffer from the problem that the results obtained depend on the choice of reference. Refinements of the basic approach are possible, where the reference is initialised and updated so as to be representative of the group of images as a whole. It is important to note, however, that even in this case the correspondence for a given image is determined solely by the information in the image and the reference. More recently, there has been considerable interest in *groupwise* methods which aim

to make more systematic use of the information in the complete set of images when establishing correspondence. The remaining two methods which were tested fall into this category.

8.1.2 Groupwise Congealing Algorithm

Learned-Miller *et al.* [70] originally introduced their 'congealing' algorithm for registering a set of hand-written digits. The aim was to avoid the arbitrary selection of a co-ordinate frame, by repeatedly registering each image with an evolving "average" model. Given the current set of transformed images (initially the original images), for each pixel position, i, the probability density function of intensities, v, at that position across the set of images, $p_i(v)$ was estimated. The objective function was then the sum of entropies of these distributions across the whole image, $F = \sum_i \int p_i(v) \log p_i(v) dv$. A set of image deformations were optimised to minimise this. In later work on registering sets of 3-D medical images [133], the objective function was approximated by $\sum_j \sum_i \log p_i(v_{ij})$, where v_{ij} is the value of pixel i in deformed image j. During optimisation, each image was warped so as to bring pixels with similar intensities into correspondence across the set. This later approach was implemented.

8.1.3 Groupwise MDL Algorithm

A groupwise method which uses a Minimum Description Length (MDL) formulation [119] was described in Chapter 4. The main idea is that the complete set of images can be encrypted as a coded message, and the description length [85] in bits used as an objective function. Rather than encoding the raw images, the encoding uses an appearance model, built using the estimated correspondences, to approximate the data; the encoding needs also to include details of the model itself and of the discrepancy between each image and its model approximation. As the registration proceeds, the correspondences, and hence the appearance model, are continually updated so as to minimise the description length.

8.2 Results of Comparison

Figure 8.1 compares the performance of the three registration algorithms outlined in Section 8.1. All the measures tested in the previous section were computed, but results are shown for only the most sensitive model-based method. Figures 8.1(a) and (d) show Specificity calculated using a shuffle radius of 2.1, for different values of m, the number of modes used to build the generative model. Figure 8.1(b) shows generalised overlap using different weightings and Figure 8.1(c) shows Generalisation. The results shown in Figures 8.1(a) and (c) suggest that the MDL groupwise approach gives the best registration result for the MGH

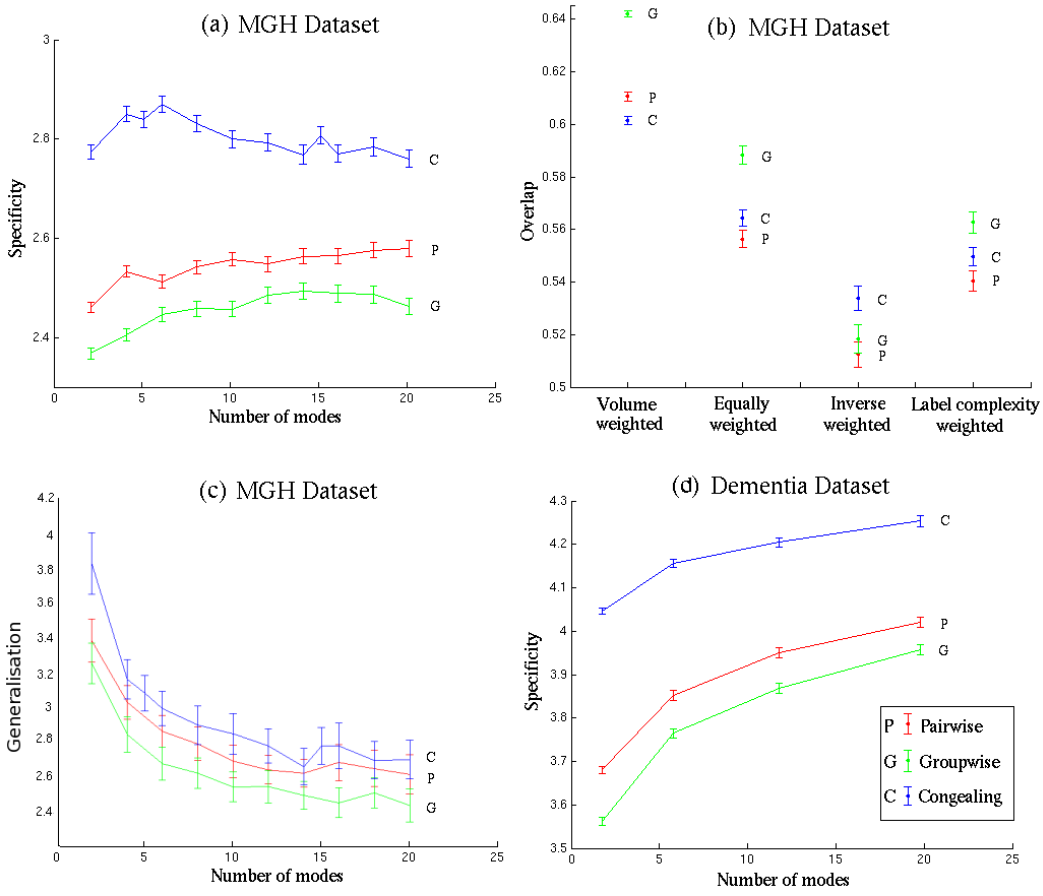

Figure 8.1: **Left and bottom:** Overlap, Generalisation and Specificity of the three registration methods as a function of the number of modes included in the appearance

Dataset, followed by Pairwise and Congealing in order of decreasing performance – irrespective of the value of m. Inspection of the error bars shows that these differences are statistically significant. The results for Generalised Overlap, shown in Figure 8.1(b), are more complicated, with the performance of the different NRR algorithms ordered differently for different weightings, though inspection of the error bars shows that many of the differences are not significant. Overall, the same general pattern emerges as for Specificity, with the Groupwise method generally best (statistically significantly in two cases), but with no significant difference between Pairwise and Congealing in most cases. The results for inverse volume weighting generally lack significance, but are inconsistent with those obtained using the other weighting schemes. Volume weighting gives the best separation between the different variants, and places the three methods in the same order as Specificity. Overall, this supports the interpretation that Specificity give results that are generally equivalent to those obtained using Generalised Overlap, but with higher sensitivity. Finally, the Specificity results shown in Figure 8.1(d) for the Dementia Dataset, place the three methods in the same order.

A single comparison, as presented in this chapter, took a few hours to complete on a modern computer. The experiments reported were performed in 2-D to limit the computational cost of running the large-scale evaluation for a range of parameter values and with repeated measurements. The extension to 3-D is, however, trivial, though the calculation of shuffle distance for 3-D images increases the computational cost sig-

nificantly. The method in 3-D has been implemented and the time taken to evaluate the registration of 100 190x190x50 images using a shuffle radius of 2.1 and $\mathcal{M} = 1000$ is around 62.5 hours on a modern PC.

8.3 Exploration of 3-D NRR

The kllkm

Chapter 9

Future Work

"We know nothing in reality; for truth lies in an abyss."

– Democritus.

THIS chapter suggests some directions for further work, providing brief details in each case of the approach that might be taken.

9.1 Automatic Model Building

The work on automatic model building has already been taken further by a team of which the author of this thesis was a member. Exploration of models and refinement of their quality by improving NRR methods led to positive results. Models of appearance were built from a triangulation of a set of images [79] and the algorithm then extended to 3-D.

9.2 Assessment of NRR

There are a few proposed modifications and extensions that can further improve the validity and accuracy of the metrics used for assessment of NRR.

9.2.1 Specificity and Generalisation

Normalisation

In the work which this thesis presents, values returned for various measures are dependent upon the size of the sets, the dimensionality and a few other free parameters. This makes it difficult to argue about and distinguish between results from different experiments, unless all conditions (free parameters) remained identical. For example, an experiment performed with large images and small sets cannot trivially be compared against other experiments involving small images and very large sets. In order for all results to be numerically comparable, there needs to be a normalisation stage, which accounts for the many free parameters simultaneously.

Another problem one can identify lies in the fact that computation of Specificity and Generalisation is not principled. Both problems can be corrected by calculating the self-normalising pseudo-entropy of graphs [74]. This graph represents the distances between images (edges) where

vertexes are individual images. Entropic graphs is an area that was explored in great depth, yet it turned out to be rather complex due to the need to estimate many parameters, using a Monte-Carlo simulation. Although efforts to adopt the method have been conceded, there is place to propose another paradigm for dealing with this issue. The *ad hoc* nature of Specificity and Generalisation leaves plenty of room for new measures to evolve. Whether an alternative method would be equally cheap to compute remains an unknown. As in many such large-scale problems, simplicity has its merits, too.

Twining and Taylor [117] addressed the problems above shortly after experimental work for this thesis had been completed. In their paper on KL divergence, a principled method is presented.

Image Separation

In Chapter 6, a measure of image distance was devised as means of measuring the overlap between distributions. With many images to be considered, one could consider systematic alterations.

Of particular interest is the notion of sensitivity as it enables one assessment method to be compared against another. Although sensitivities were shown to culminate at a particular value of the shuffle distance, other approaches that had been investigated could perhaps entail superior sensitivity, at the expense of computational power. It is worth exploring and finding if a more complex approach, e.g. one which considers

an average or median in a neighbourhood of pixels/voxels, outperforms shuffle distance.

9.2.2 Investigating Robustness

One aspect which must never be neglected are the boundaries and edge cases. It is valuable to know where the methods cease to be usable; for example, when noise levels supersede the signal. Having identified the limitations of the method, their robustness can be improved. For instance, in the case of NRR assessment, one can improve the range of displacements where results can be differentiated by increasing the size of the shuffle neighbourhood. Prior experiments showed that the performance is degraded beyond a certain shuffle neighbourhood size, but there are other parameters that can be varied, so their effect on the shuffle neighbourhood is not mutually exclusive.

9.2.3 Extension to 3-D

Computational Cost

Computational loads are an important factor that has become a barrier, particularly in 3-D – an extension that is described was alluded to, but not described in detail. The model and NRR assessment methods have already been extended to operate on three-dimensional data. Rather than

handling 2-D images, the methods then deal with volumes and, in accordance to this, shuffle distance neighbourhoods become a box or sphere of voxels, rather than a square or a disk.

There are particular steps in the algorithm whose computational cost is far greater than the remainder. Firstly, one must consider the long time that is required to synthesise many images from appearance models and subsequently use them in an evaluation. The greater the number of synthetic images, the more accurate the results. This relationship means that there is no clear point of balance.

Secondly, the more time-consuming process involves the computation of inter-image distances. With the added complexity of a third dimension, as well as a shuffle distance with large neighbourhood sizes, there is a considerable cost which is proportional to the number of voxels at hand. If image size is increased, it is increased in many dimensions, which slows down the heart of computation – measuring image distances. At present, speed is relatively satisfactory, but there is much left to be desired, especially because of the scalability of the problem. The previous chapter provided insight into the duration of such experiments, which scale linearly with respect to the *depth* of an image volume.

Methods of Speedup

There are several tricks-of-the-trade, which can be used to speed up the process of registration assessment and model evaluation in 3-D. A few

Figure 9.1: A multi-resolution approach illustrated in 2-D. Coarser representations are shown at the top levels and the original image lies at the bottom.

examples follow.

Multi-scale. One strategy, for example, involves resampling the data, then dealing with smaller (downsized) versions of the whole, as shown in Figure 9.1. The basic concept is that, if the scale of the problem is reduced, it can be handled at a coarse level and then iteratively handled at finer levels, until the original unscaled data is reached.

Progressively-improved estimates. Another strategy worth employing and exploring is one which uses a figure of merit that is gradually made more accurate. In the case where synthetic images are considered, this strategy becomes a possibility. The size of the synthetic set can be simply limited. However, an assessment method

that is quicker to deliver an estimate is one which makes use of the entire set (whether coarse or not) rather than use just a subset that is subjected to bias.

There are different ways of reducing the complexity of the problem at hand. In practice, this means that a box of voxel may be sliced into 8 (2^3) equal-sized boxes which are then used in the analysis, or even more usefully, the box should be rescaled to become 8 times smaller, in terms of volume.

Selective assessment of slices. In general, there is not much which distinguishes the method's use in 2-D and in 3-D, other than efficiency factors. However, several possibilities emerge owing to the fact that 3-D data can be interpreted differently once its dimensionality is reduced. For example, one can choose one representative slice from a larger volume and refine the evaluation by considering more slices, one at a time. This suffers from the fact that voxels whose position varies in the third dimension, i.e. it moves between the slices due to warping, will not be treated appropriately. All these issues, along with other pitfalls, will be addressed in the near future.

Chapter 10

Summary and Conclusions

"If you're not part of the solution, you're part of the precipitate."

– *Henry J. Tillman.*

THE work covered in this thesis can be summarised as follows. Firstly, a novel framework was described which non-rigidly registers images using a model-based similarity measure. This framework is able to deal with any type of images and, while it requires a a sufficiently large number of images in order to become practical (i.e. in order for a sensible model to be built), its performance does not depend on the type of variation that is contained in the set of images. As a result of registering the images using a model-based approach, one also obtains an appearance model, which is progressively refined and whose quality is dependent on the quality of the registration algorithm. This establishes a framework

for automatic construction of models that requires a model-based objective function.

The second part of the work is concerned with assessment. Two things are being assessed: the quality of any appearance model (or any generative model) and the quality of a groupwise registration. This opens up the possibilities of comparing NRR algorithms and supports evidence-based development of improved methods.

10.1 Discussion

The results of the validation experiments reported in Chapter 7 are the most important outcome of the work presented here. They demonstrate a causal relationship between our Specificity and Generalisation measures, and a known (up to an additive constant) mean pixel displacement, d. A strong correlation between these model-based measures and a Generalised Overlap measure, based on ground truth, adds further weight to this interpretation. The fact that the relationship with d held good over many different instantiations of a very general class of perturbing warps, makes it unlikely (though not impossible) that there is any significant dependence on variation patterns.

The results obtained with added noise are also encouraging, since it is a reasonable concern that the use of an intensity-based distance measure might make the model-based measures sensitive to noise. In the event,

the approach seems robust to quite significant levels of noise. The fact that the absolute values of specificity and generalisation change when noise is added, mean that they would not be useful for comparing registration results for different image sets. Their ability to compare the performance of different registration algorithms applied to the same set of images, the main intended use, is, however, unaffected.

Our results comparing the performance of different registration algorithms demonstrate that the model-based measures, and Specificity in particular, are sufficiently sensitive to misregistration to provide useful discrimination in a practical setting. There is, however, a potential concern that is important to address. It might be argued that using a model-based approach to assessing registration favours methods which use a model-based objective function for registration (as in the experiments reported here). In practice, there is no reason to believe that this is a problem.

First, as argued above, the validation results show that there is a causal relationship between the mean pixel displacement, d, and Specificity or Generalisation. It is thus irrelevant how a registration (or misregistration) has been obtained. Second, the MDL objective function optimised in the model-based registration method measures a quite different property of the model to those used in evaluation, so there is no element of 'self-fulfilling prophecy'. In an ideal world it would, of course, be preferable to avoid even the possibility of bias, though it seems unlikely that one

could devise a strategy for evaluation that had no relevance to achieving a good registration in the first place. In due course, other ground-truth-free methods of evaluation may be developed, allowing a multi-perspective assessment of performance.

One obvious limitation of this approach to evaluation is that it can *only* be applied to groups of images. This could be considered an important restriction, since many practical applications involve registration of pairs or very short temporal sequences of images. One would argue that, in fact, this is a necessary restriction, because it is only possible to arrive at a meaningful assessment of registration in the context of a population of images.

There are a number of issues that merit further investigation. A particular method of measuring image separation has been studied, but others, such as local correlation, would be worth exploring. Another interesting issue is whether it is possible within this framework to localise registration errors. Some initial experiments have been performed, summing the shuffle difference maps between all pairs of images in the registered set. This gives some interesting results, highlighting areas of common misregistration, but it is not clear what quantitative interpretation could be placed on such maps. Finally, it is clear that the current measures of Specificity and Generalisation are not normalised – their values depend on the size of the set of registered images, the number of synthetic images generated and so on [117]. It is clearly worth exploring the possibility

of measuring more fundamental properties of the relationship between the real and synthetic image distributions, with a view to achieving a 'natural' normalisation as discussed in Chapter 9.

10.2 Conclusions

A method for registering images in a groupwise fashion has been described, including the quality of their appearance model in the objective function. Not only does this enable an algorithm to reach good results from a groupwise perspective, but it also results in the automated construction of appearance model, whose quality is assessed.

A model-based approach to evaluating the results of NRR of a group of images has also been described. The most important advantage of the new method is that it does not require any ground truth, but depends solely on the registered images themselves.

The approach has been validated by studying the effect of perturbing, progressively, the registration of an initially registered set of images, comparing the results with those obtained using a 'gold standard' measure based on the overlap of ground-truth anatomical labels. It has been shown that the new method provides measures of registration accuracy that are monotonic functions of the known misregistration, and that one, *Specificity*, provides a more sensitive measure of misregistration than the approach based on ground truth.

The model-based approach requires a distance measure in image space, and it has also been demonstrated that the use of shuffle distance, rather than Euclidean distance, improves the sensitivity of the approach.

A set of experiments have further validated the approach, and illustrated its application, by performing a comparative evaluation of the results obtained using three different NRR algorithms, demonstrating the superiority of a fully-groupwise algorithm over a repeated pairwise approach.

It is important to emphasise that this approach is not restricted to evaluating model-based NRR algorithms, though the thesis presented results for one such method; the model-based measures of registration accuracy can be applied to any set of non-rigidly registered images, however they were obtained. The possibility of a bias in favour of model-based methods of registration has been discussed. It was concluded that there is no major problem, though it would be desirable to compare results obtained using a range of ground-truth-free methods of evaluation.

Bibliography

[1] A. Abkar and H. Hedenmalm, “A Riesz Representation Formula for Super-Biharmonic Functions,” Annales Academie Scientiarum Fennice Mathematica, vol. 26, pp. 305–324, 2001.

[2] J. Ahlberg and R. Forchheimer, “Face tracking for model-based coding and face animation,” International Journal of Imaging Systems and Technology, vol. 13, pp. 8–22, 2003.

[3] J. Ashburner and K. J. Friston, “Why Voxel-Based Morphometry Should Be Used,” NeuroImage, vol. 14, pp. 1238–1243, 2001.

[4] M. F. Beg, M. I. Miller, T. A., and L. Younes, “Computational Anatomy: Computing Metrics on Anatomical Shapes," presented at Proceedings of the First IEEE International Symposiumon Biomedical Imaging, 2002.

[5] S. Baker, I. Matthews, and J. Schneider, “Automatic construction of active appearance models as an image coding problem,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, pp. 1380–1384, 2004.

[6] C. J. Beeston and C. J. Taylor, “Automatic landmarking of cortical sulci,” in Medical Image Computing and Computer-Assisted Intervention, Lecture Notes in Computer Science, vol. 1935, pp. 125–133, 2000.

[7] M. Beauchemin and K. P. B. Thomson, “The evaluation of segmentation results and the overlapping area matrix,” International Journal of Remote Sensing, 18(18):3895–3899, 1997.

[8] D. Beymer and T. Poggio. “Image Representations for Visual Learning,” in Science, vol. 272, issue 5270, pp. 1905–1909.

[9] K. K. Bhatia, J. V. Hajnal, B. K. Puri, A. D. Edwards, and D. Rueckert, “Consistent groupwise non-rigid registration for atlas construction,” presented in IEEE International Symposium on Biomedical Imaging, Arlington, VA, USA, 2004.

[10] F. L. Bookstein, “Principal Warps: Thin-Plate Spines and the Decomposition of Deformations,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 11, pp. 567–585, 1989.

[11] F. L. Bookstein, “Linear Methods for Nonlinear Maps: Procustes Fits, Thin-Plate Splines and the Biometric Analysis of Shape Variability," in Brain Warping, A. W. Toga, Ed. San Diego, CA, USA: Academic Press, pp. 157–181, 1999.

[12] A. D. Brett and C. J. Taylor, “A method of automated landmark generation for automated 3D PDM construction,” Image and Vision Computing, vol. 18, pp. 739–748, 2000.

[13] P. Cachier and D. Rey, “Symmetrization of the Non-Rigid Registration Problem Using Inversion-Invariant Energies: Application to Multiple Sclerosis," presented at Third International Conference on Medical Image Computing and Computer Assisted Intervention, 2000.

[14] V. Camion and L. Younes, “Geodesic Interpolating Splines," presented at Proceedings of Energy Minimisation in Computer Vision and Pattern Recognition, 2001.

[15] G. E. Christensen, R. D. Rabbitt, and M. I. Miller, “Deformable templates using large deformation kinematics,” in IEEE Transactions on Image Processing, vol. 5:1010, pp. 1435–1447, 1996.

[16] W. R. Crum, O. Camara, D. Rueckert, K. Bhatia, M. Jenkinson, and D. L. G. Hill, “Generalised overlap measures for assessment of pairwise and groupwise image registration and segmentation,” in Proceedings of Medical Image Computing and Computer-Assisted Intervention, Lecture Notes in Computer Science, vol. 3749. pp. 99–106, 2005.

[17] W. R. Crum, T Hartkens, and D. L. G. Hill, “Non-rigid image registration: theory and practice,” British Journal of Radiology, vol. 77, pp. 140–153, 2004.

[18] T. F. Cootes, C. Beeston, G. J. Edwards, and C. J. Taylor, “A unified framework for atlas matching using Active Appearance Models,” in Information Processing in Medical Imaging, Proceedings, vol. 1613, Lecture Notes in Computer Science, 1999, pp. 322–333.

[19] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, pp. 681–685, 2001.

[20] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” in Proceedings of the European Conference on Computer Vision (ECCV), Lecture Notes in Computer Science, vol. 1407. Springer, pp. 484–498, 1998.

[21] T. F. Cootes, S. Marsland, C.J. Twining, K. Smith, and C.J. Taylor, “Groupwise diffeomorphic non-rigid registration for automatic model building,” in Proceedings of the European Conference of Computer Vision 2004, pp. 316–32.

[22] T. F. Cootes, C. J. Twining, V. Petrovic, R. Schestowitz, and C. J. Taylor, "Groupwise Construction of Appearance Models using Piece-wise Affine Deformations." in Proceedings of the British Machine Vision Conference (BMVC'04), Kingston UK, 2004.

[23] T. F. Cootes, G. V. Wheeler, K. N. Walker, and C. J. Taylor, "View-based active appearance models," Image and Vision Computing, vol. 20, pp. 657–664, 2002.

[24] T. F. Cootes, G. V. Wheeler, K. N. Walker, and C. J. Taylor, "Coupled-view active appearance models," British Machine Vision Conference, vol. 1, pp. 52–61, 2000.

[25] R. H. Davies, C. J. Twining, T. F. Cootes, J. C. Waterton, and C. J. Taylor, "3D statistical shape models using direct optimisation of description length," in Proceedings of the European Conference on Computer Vision, Lecture Notes in Computer Science, vol. 2352, pp. 3–20, 2002.

[26] R. H. Davies, C. J. Twining, T. F. Cootes, J. C. Waterton, and C. J. Taylor, "minimum description length approach to statistical shape modeling," IEEE Transactions on Medical Imaging, vol. 21, pp. 525–537, 2002.

[27] R. H. Davies, C. J. Twining, P. D. Allen, T. F. Cootes, and C. J. Taylor, "Shape discrimination in the hippocampus using an MDL model," in Information Processing in Medical Imaging Proceedings, Lecture Notes in Computer Science, vol. 2732, pp. 38–50, 2003.

[28] B. M. Dawant, "Non-Rigid Registration of Medical Images: Purpose and Methods, A Short Survey," in Proceedings of the First IEEE International Symposium on Biomedical Imaging, 2002.

[29] S. Duchesne, J. C. Pruessner, and D. L. Collins, "Appearance-based segmentation of medial temporal lobe structures," NeuroImage, vol. 17, pp. 515–531, 2002.

[30] P. Dupuis, U. Grenander, and M. I. Miller, "Variational problems on flows of diffeomorphisms for image matching," Quarterly of Applied Mathematics, vol. 56, pp. 587 – 600, 1998.

[31] G. J. Edwards, C. J. Taylor, and T. F. Cootes, "Interpreting face images using active appearance models," in IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998.

[32] G. J. Edwards, T. F. Cootes, and C. J. Taylor, "Face recognition using active appearance models," in Proceedings of European Conference on Computer Vision, Lecture Notes in Computer Science, vol. 2. pp. 581–595, 1998.

[33] P. F. Felzenszwalb, "Representation and detection of deformable shapes," Computer Vision and Pattern Recognition, vol. 1, pp. 102–108, 2003.

[34] A. F. Frangi, D. Rueckert, J. A. Schnabel, and W. J. Niessen, "Automatic construction of multiple-object three-dimensional statistical shape models: application to cardiac modelling," IEEE Transations in Medical Imaging, vol. 21, pp. 1151–1166, 2002.

[35] J. Gee, M. Reivich, and R. Bajcsy, “Elastically deforming 3D atlas to match anatomical brain images," Journal of Computer Assisted Tomography, vol. 17, pp. 225-236, 1993.

[36] F. Glover, E. D. Taillard, and D. de Werra, “A user’s guide to taboo search,” Annals of Operations Search, pp. 3–28, 1993.

[37] U. Grenander and M. I. Miller, “Computation Anatomy: An Emerging Discipline," in Quarterly of Applied Mathematics, vol. 56, 1998, pp. 617 – 694.

[38] H. C. Grunau and G. Sweers, “Maximum principles and positive principal eigenfunctions for polyharmonic equations," presented at Reaction Diffusion Systems, 1998.

[39] A. Guimond, J. Meunier, and J. Thirion, “Automatic computation of average brain models,” presented at Medical Image Computing and Computer Assisted Intervention 1998, pp. 631–640, Cambridge, MA, USA.

[40] A. Guimond, J. Meunier, and J.-P. Thirion, “Average Brain Models: A Convergence Study," INRIA, Sophia Antipolis 1999.

[41] J. V. Hajnal, D. L. G. Hill, and D. J. Hawkes, “Medical image registration,” Boca Raton, Fla. ; London: CRC Press, 2001.

[42] S. Haker, A. Tannenbaum, and R. Kikinis, “Mass Preserving Mappings and Image Registration,” presented at Fourth International Conference on Medical Image Computing and Computer Assisted Intervention, 2001.

[43] P. F. Hemler, P. A. van den Elsen, T. Sumanaweera, S. Napel, J. Drace, and J. R. Adler, “A quantitative comparison of residual errors for three different multimodality registration techniques,” in Information Processing in Medical Imaging, pp. 251–262, 1995.

[44] S. Henn and K. Witsch, “Iterative Multigrid Regularisation Techniques for Image Matching," SIAM Journal of Scientific Computing, vol. 23, pp. 1077–1093, 2001.

[45] D. W. Hansen, M. Nielsen, J. P. Hansen, A. S. Johansen and M. B. Stegmann, “Tracking eyes using shape and appearance,” IAPR Workshop on Machine Vision Applications, pp. 201–204, 2002.

[46] P. Hellier, C. Barillot, I. Corouge, B. Giraud, G. L. Goualher, L. Collins, A. Evans, G. Malandain, and N. Ayache, “Retrospective evaluation of inter-subject brain registration,” in Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI), Lecture Notes in Computer Science, vol. 2208. Springer, 2001, pp. 258–265.

[47] A. Hill, C. J. Taylor, and A. D. Brett, “A framework for automatic landmark identification using a new method of nonrigid correspondence,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, pp. 241–251, 2000.

[48] D. L. G. Hill, D. J. Hawkes, and J. E. Crossman, “Registration of MR and CT images for skull base surgery using point-like anatomical features,” in British Journal of Radiology, vol. 64, pp. 1330–1035, 1991.

[49] D. R. Jenkins, “Thin Plate Spline Interpolation on an Annulus," presented at Proceedings of the 1999 International conference on Computational Techniques and Applications, 2000.

[50] T. Jebara, “Images as Bags of Pixels,” presented at International Conference of Computer Vision, Proceedings, Nice, France, 2003.

[51] I.T. Joliffe, “Principal Component Analysis,” Springer Series in Statistics, Springer, New York, 1986.

[52] S. C. Joshi and M. L. Miller, “Landmark Matching via Large Deformation Diffeomorphisms,” IEEE Transaction on Image Processing, vol. 9, pp. 1357–1370, 2000.

[53] J. F. Kenney and E. S. Keeping, “Mathematics of Statistics,” part 2, 2nd edition, Princeton, New Jersey: Van Nostrand, 1951.

[54] A. C. W. Kotcheff and C. J. Taylor, “Automatic construction of eigenshape models by genetic algorithm,” in Information Processing in Medical Imaging, vol. 1230, pp. 1–14, Lecture Notes in Computer Science, 1997. 2005.

[55] K. N. Kutulakos, “Approximate n-view stereo,” in Proceedings of the European Conference on Computer Vision (ECCV), Lecture Notes in Computer Science, vol. 1842. Springer, 2000, pp. 67–83.

[56] A. Lanitis, “PROSOPO - A face image synthesis system,” Advances in Informatics, Lecture Notes in Computer Science, vol. 2563, pp. 297–315, 2003.

[57] A. Lanitis, C. J. Taylor, and T. F. Cootes, “Automatic Face Identification System Using Flexible Appearance Models,” Image and Vision Computing, vol. 13, pp. 393–401, 1995.

[58] A. Lanitis, C. J. Taylor, and T. F. Cootes, “Toward automatic simulation of aging effects on face images,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, pp. 442–455, 2002.

[59] L. LeBriquer and J. Gee, “Design of a statistical model of brain shape,” presented at Proceedings of IPMI, 1997.

[60] S. Lee, G. Wolberg, and S. Y. Shin, “Scattered Data Interpolation with Multilevel B-Splines," IEEE Transactions on Visualisation and Computer Graphics, vol. 3, pp. 228 – 244, 1997.

[61] Y. Li, S. Gong, and H. Liddel, “Constructing facial identity surfaces,” in Computer Vision and Pattern Recognition, 2001.

[62] B. Likar and F. Pernus, “A Hierarchical Approach to Elastic Registration Based on Mutual Information,” Image and Vision Computing, vol. 19, pp.

[63] J. A. Little, D. L. G. Hill, and D. J. Hawkes, “Deformations Incorporating Rigid Structures,” in IEEE Workshop on Mathematical Methods in Biomedical Image Analysis, pp. 104–113, 1996.

[64] J. Lötjönen and T. Mäkelä, “Elastic Matching Using a Deformation Sphere,” presented at Fourth International Conference on Medical Image Computing and Computer Assisted Intervention, 2001.

[65] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, “Multimodality image registration by maximization of mutual information,” IEEE Transactions on Medical Imaging, vol. 16, issue 2, pp. 187–198, 1997.

[66] J. B. A. Maintz and P. A. van den Elsen and M. A. Viergever, “Comparison of edge-based and ridge-based registration of CT and MR brain images,” Medical Image Analysis, vol. 1:2, pp. 151–161, 1996.

[67] D. Marr, “Understanding complex information processing systems,” Vision, pp. 19–24, 1982.

[68] S. Marsland, C. J. Twining, and C. J. Taylor, “Groupwise non-rigid registration using polyharmonic clamped-plate splines,” presented at Medical Image Computing and Computer-Assisted Intervention, Montreal, Canada, 2003.

[69] C. R. Maurer, G. B. Aboutanos, B. M. Dawant, R. J. Maciunas, and J. M. Fitzpatrick, “Registration of 3-D Images Using Weighted Geometrical Features,” IEEE Transaction on Medical Imaging, vol. 15:6, pp. 836–849, 1996.

[70] E. G. Miller, N. E. Matsakis, and P. A. Viola, “Learning from one example through shared densities on transforms,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Volume 1, 2000, pp. 464–471.

[71] M. I. Miller, S. C. Joshi, and G. E. Christensen, “Large Deformation Fluid Diffeomorphisms for Landmark and Image Matching," in Brain Warping, A. W. Toga, Ed. San Diego, CA, USA: Academic Press, 1999, pp. 115 – 131.

[72] T. Mitchell, “Machine Learning,” New York ; London: McGraw-Hill, 1997.

[73] A. Najmi, R. A. Olshen, and R. M. Gray, “A criterion for model selection using minimum description length,” in Proceedings Compression and Complexity of Sequences, pp. 204–214, Salerno, Italy, 1997.

[74] H. Neemuchwala, A. O. Hero, and P. Carson, “Image registration using entropy measures and entropic graphs,” European Journal of Signal Processing, 2003.

[75] W. Peckar, C. Schnörr, K. Rohr, and H. S. Stiehl, “Non-Rigid Image Registration using a Parameter-Free Elastic Model," presented at Proceedings of the Ninth British Machine Vision Conference (BMVC’98), 1998.

[76] C. A. Pelizzari, G. T. Y. Chen, D. R. Spelbring, R. R. Weichselbaum, and C. Chen, "Accurate three dimensional registration of CT, PET and/or MR images of the brain," Journal of Computer Assisted Tomography, vol. 13, pp. 2–26, 1989.

[77] G. P. Penney, J. A. Little, J. Weese, D. L. G. Hill, D. J. Hawkes, "A comparison of similarity measures for use in 2D-3D medical image registration," IEEE Transaction on Medical Imaging, vol. 17, pp. 586–595, 1998.

[78] M. Petrou and P. Bosdogianni, "Image processing: the fundamentals," ISBN 0471 99883 4, 1999.

[79] V. S. Petrovic, T. F. Cootes, C. J. Twining, and C. J. Taylor, "Automatic Framework for Medical Image Registration, Segmentation and Modeling," Proceedings of Medical Image Understanding and Analysis. Vol. 2, pp.141–145, 2006.

[80] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Image Registration by Maximisation of Combined Mutual Information and Gradient Information," IEEE Transaction on Medical Imaging, vol. 19, pp. 809–814, 2000.

[81] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual information matching in multiresolution contexts," in Image and Vision Computing, vol. 19, issues 1–2, pp. 45–52, January 2001.

[82] I. Praktikis, C. Barillot, and P. Hellier, "Robust Multi-Scale Non-Rigid Registration of 3D Ultrasound Images," presented at Scale-Space 2001, 2001.

[83] W. H. Press, "Numerical recipes in C/C++ : the art of scientific computing." Cambridge: Cambridge University Press, 2002.

[84] M. Rabbani and R. Joshi, "An overview of the JPEG 2000 still image compression standard," Signal Processing: Image Communication, vol. 17, pp. 3–48.

[85] J. R. Rissanen, "Stochastic complexity in statistical inquiry," presented at World Scientific Series in Computer Science, Singapore, 1989.

[86] P. Rogelj and S. Kovacic, "Similarity Measures for Non-Rigid Registration," presented at Medical Imaging 2001: Image Processing.

[87] P. Rogelj, S. Kovacic, and J. C. Gee, "'Validation of a nonrigid registration algorithm for multimodal data," in Proceedings of Medical Imaging 2002, Image Processing, SPIE Proceedings, vol. 4684, 2002, pp. 299–307.

[88] T. Rohlfing and J. Maurer, Calvin R., "Intensity-Based Non-Rigid Registration Using Adaptive Multilevel Free-Form Deformation with an Incompressibility Constraint," presented at Fourth International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'01), 2001.

[89] K. Rohr, "Landmark-Based Image Analysis Using Geometric and Intensity Models," Kluwer, 2001.

[90] S. Romdhani, S. Gong, and A. Psarrou, “A multi-view nonlinear active shape model using kernel PCA.” in Proceedings of the British Machine Vision Conference, pp. 483-492, 1999.

[91] D. Rueckert, A. F. Frangi, and J. A. Schnabel, “Automatic construction of 3-D statistical deformation,” IEEE Transactions on Medical Imaging, vol. 22, issue 8, pp. 1014–1025, 2003.

[92] D. Rueckert, A. F. Frangi and J. A. Schnabel, “Automatic construction of 3D statistical deformation models using non-rigid registration,” presented at Medical Image Computing and Computer-Assisted Intervention 2001.

[93] R. S. Schestowitz, C. J. Twining, T. F. Cootes, V. S. Petrovic, C. J. Taylor, and B. Crum, “Assessing the Accuracy of Non-Rigid Registration With and Without Ground Truth,” IEEE International Symposium on Biomedical Imaging (ISBI), 2006.

[94] R. S. Schestowitz, C. J. Twining, T. F. Cootes, and C. J. Taylor, “Image Registration by Model Criteria,” presented in Proceedings of MIAS-IRC Plenary Meeting, pp. 16–17, 2004.

[95] R. S. Schestowitz, C. J. Twining, T. F. Cootes, V. S. Petrovic, and C. J. Taylor, “A Generic Method for Evaluating Appearance Models,” presented in Proceedings of MIAS-IRC Plenary Meeting, 2006.

[96] J. A. Schnabel, C. Tanner, A. Castellano-Smith, A. Degenhard, M. O. Leach, D. R. Hose, D. L. G. Hill, and D. J. Hawkes, “Validation of non-rigid image registration using finite element methods: application to breast MR images,” IEEE Transactions on Medical Imaging, vol. 22.

[97] I. M. Scott, T. F. Cootes, and C. J. Taylor, “Improving appearance model matching using local image structure," in Proceedings of Information Processing in Medical Imaging, Lecture Notes in Computer Science, vol. 2732, pp. 258–269, 2003.

[98] D. W. Shattuck, S. R. Sandor-Leahy, K. A. Schaper, D. A. Rottenberg, and R. M. Leahy, “Magnetic resonance image tissue classification using a partial volume model,” NeuroImage, vol. 13, pp. 856–876, 2001.

[99] Y. Shi and F. Qi, “Feature-based Deformable Registration of Brain Images,” 17th IEEE Symposium on Computer-Based Medical Systems (CBMS), pp. 553, 2004.

[100] M. Sonka, V. Hlavac, and R. Boyle, “Image processing, analysis and machine vision,” Pacific Grove, Calif. ; London: PWS Publishing, 1999.

[101] M. B. Stegmann, B. K. Ersboll, and R. Larsen, “FAME - A flexible appearance modeling environment,” IEEE Transactions on Medical Imaging, vol. 22, pp. 1319–1331, 2003.

[102] M. B. Stegmann, “Analysis of 4D cardiac magnetic resonance images,” Journal of The Danish Optical Society, vol. 4, pp. 38–39, 2001.

[103] S. C. Strother, J. R. Anderson, X. L. Xu, J. S. Liow, J. S., D. C. Bonar, and D. A. Rottenberg, “Quantitative comparisons of image registration techniques based on high-resolution MRI of the brain,” Journal of Computer Assisted Tomography, vol. 18, pp. 954–962, 1994.

[104] C. Studholme, D. L. G. Hill, and D. J. Hawkes, “An overlap invariant entropy measure of 3D medical image alignment,” Pattern Recognition, vol. 32, issue 1, pp. 71–86, 1999.

[105] C. Studholme, D. L. G. Hill, and D. J. Hawkes, “Multiresolution Voxel Similarity Measures for MR-PET registration,” Information Processing in Medical Imaging, pp. 287–298, 1995.

[106] N. A. Thacker, D. Prendergast, and P. I. Rockett, “B-fitting: an estimation technique with automatic parameter selection,” presented at British Machine Vision Conference 1996.

[107] P. Thevenaz, U. E. Ruttimann, and M. Unser, “A Pyramid Approach to Subpixel Registration Based on Intensity,” IEEE Transactions on Image Processing, vol. 7, pp. 27–41,1998.

[108] J. P. Thirion, “Fast Non-Rigid Matching of 3D Medical Images,” INRIA, number 2547, 1995

[109] P. M. Thompson, M. S. Mega, C. Vidal, J. L. Rapoport, and A. W. Toga, “Detecting Disease-Specific Patterns of Brain Structure using Cortical Pattern Matching and a Population-Based Probabilistic Brain Atlas,” presented at Proceedings of the 17th International Conference on Information Processing in Medical Imaging, 2001.

[110] P. Thompson and A. W. Toga, “A Surface-Based Technique for Warping Three-Dimensional Images of the Brain,” IEEE Transactions on Medical Imaging. vol. 15, pp. 402–417, 1996.

[111] A. Toga, “Brain Warping,” San Diego, CA, USA: Academic Press, 1999.

[112] A. Trouve, “Diffeomorphisms Groups and Pattern Matching in Image Analysis,” International Journal of Computer Vision, vol. 28, pp. 213–221, 1998.

[113] A. Trouve and L. Younes, “Diffeomorphic Matching Problems in One Dimension: Designing and Minimizing Matching Functionals,'' presented at Proceedings of European Conference on Computer Vision, 2000.

[114] C. J. Twining and S. Marsland, “Constructing diffeomorphic representations of non-rigid registrations of medical images,” presented at Information Processing in Medical Imaging 2003.

[115] C. J. Twining, S. Marsland, and C. J. Taylor, "Groupwise non-rigid registration: the minimum description length approach," British Machine Vision Conference 2004.

[116] C. J. Twining and C. J. Taylor, "The use of kernel principal component analysis to model data distributions," Pattern Recognition, vol. 36, pp. 217–227, 2003.

[117] C. J. Twining and C. J. Taylor, "Specificity as a Graph-Based Estimator of Cross-Entropy," in Proceedings of the British Machine Vision Conference, vol. 2, pp. 459–468, 2006.

[118] C. J. Twining, S. Marsland, and C. J. Taylor, "Measuring geodesic distances on the space of bounded diffeomorphims," in Proceedings of the British Machine Vision Conference (BMVC'02), 2002.

[119] C. J. Twining, T. F. Cootes, S. Marsland, V. Petrovic, R. Schestowitz, and C. J. Taylor, "A unified information-theoretic approach to groupwise non-rigid registration and model building," in Proceedings of Information Processing in Medical Imaging (IPMI), Lecture Notes in Computer Science, vol. 3565. Springer, 2005, pp. 1–14.

[120] P. Viola and W. M. Wells, "Alignment by maximization of mutual information," International Journal of Computer Vision, vol. 24, pp. 137–154, 1997.

[121] K. N. Walker, T. F. Cootes, and C. J. Taylor, "Determining correspondences for statistical models of appearance," in Computer Vision - ECCV 2000, Pt I, Proceedings, vol. 1842, Lecture Notes in Computer Science, 2000, pp. 829–843.

[122] Y. Wang and L. H. Staib, "Elastic model based non-rigid registration incorporating statistical shape information," presented at Medical Image Computing and Computer-Assisted Intervention, 1998.

[123] L. Wang, Y. Zhang, and J. Feng, "On the euclidean distance of images," IEEE Trans. Pattern Anal. Machine Intell., vol. 27, pp. 1334–1339,

[124] Y. Wang and L. H. Staib, "Integrated approaches to non-rigid registration in medical images," presented at Workshop on Applications of Computer Vision, 1998.

[125] S. K. Warfield, J. Rexilius, P. S. Huppi, T. E. Inder, E. G. Miller, W. M. Wells, III, G. P. Zientara, F. A. Jolesz, and R. Kikinis, "An Entropy Measure to Assess Nonrigid Registration Algorithms for Statistical Atlas Construction," presented at Medical Image Computing and Computer-Assisted Intervention, 2001.

[126] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation," IEEE Transactions on Medical Imaging. vol. 23, pp. 903–21, 2004.

[127] J. West, J. M. Fitzpatrick, M. Y. Wang, B. M. Dawant, J. Maurer, C. R., R. M. Kessler, R. J. Maciunas, C. Barillot, D. Lemoine, A. Collignon, F. Maes, P. Suetens, D. Vandermeulen, P. A. van den Elsen, S. Napel, T. S. Sumanaweera, B. Harkness, P. F. Hemler, D. L. G. Hill, D. J. Hawkes, C. Studholme, J. B. A. Maintz, M. A. Viergever, G. Malandain, X. Pennec, M. E. Noz, J. Maguire, G. Q., M. Pollack, C. A. Pelizzari, R. A. Robb, D. Hanson, and R. P. Woods, "Comparison and evaluation of retrospective intermodality brain image registration techniques," Journal of Computer Assisted Tomography, vol. 21, pp. 554–566, 1997.

[128] M. Xu and W. L. Nowinski, "Talairach-Tournoux Brain Atlas Registration Using a Metalforming Principle-Based Finite Element Method," Medical Image Analysis, vol. 5, pp. 271–279, 2001.

[129] C. Yam, M. S. Nixon, and J. N. Carter, "Automated person recognition by walking and running via model-based approaches," Pattern Recognition, vol. 37, 2004.

[130] L. Younes, "Computable Elastic Distances Between Shapes," SIAM Journal of Applied Mathematics, vol. 58, pp. 565 – 586, 1998.

[131] L. Younes, "Deformations, Warping and Object Comparison: A Tutorial," 2000.

[132] J. Zhang and A. Rangarajan, "A unified feature-based registration method for multimodality images," IEEE International Symposium on Biomedical Imaging (ISBI), vol. 1, pp. 724-727, 2004.

[133] L. Zollei, E. Learned-Miller, E. Grimson, and W. Wells, "Efficient population registration of 3D data," in Workshop on Computer Vision for Biomedical Image Applications: International Conference of Computer Vision (ICCV05), 2005.

[134] B. Zitova and J. Flusser, "Image registration methods: A survey," Image and Vision Computing, vol. 21, pp. 977–1000, 2003.