Optimisation Methods for Medical Image Registration

Stefan Klein

Optimisation Methods for Medical Image Registration
Stefan Klein
PhD thesis, Utrecht University, the Netherlands.

The research described in this thesis was carried out at the Image Sciences Institute, University Medical Center Utrecht, Utrecht, the Netherlands. The project was financially supported by the Netherlands Organisation for Scientific Research (NWO). Grants for publication of this thesis have been kindly provided by the Röntgen Stichting Utrecht, the research school ImagO, Nucletron BV, and 3mensio Medical Imaging BV.

ISBN 978-90-8891-064-7
Printed by Uitgeverij BOX Press, the Netherlands.

Optimisation Methods for Medical Image Registration

Optimalisatiemethoden voor het Registreren
van Medische Beelden
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof.dr. J.C. Stoof, ingevolge het
besluit van het college voor promoties in het openbaar te verdedigen
op donderdag 9 oktober 2008 des middags te 4.00 uur

door

Stefan Klein

geboren op 20 april 1978 te Almelo.

Promotor: Prof. dr. ir. M.A. Viergever

Co-promotor: Dr. J.P.W. Pluim

Preface

Going through a PhD trajectory is a bit like solving an optimisation problem. It always takes more time than you expect. You can go in the wrong direction, get stuck for a while in a local dip, or suddenly make big progress. Moreover, hard constraints (deadlines) are often necessary to ensure convergence. A smooth trajectory towards the end point can only be realised with the help of other people. Gratitude goes towards them. They make the difference between an *ill-posed* and a *well-posed* optimisation problem.

First of all, I thank Max Viergever. Being the promotor, you are the start and end condition of a PhD study. More than that, I appreciate the valuable input during our meetings, and your enormous drive to "transform" research into publications. Second, I want to thank my co-promotor and direct supervisor: Josien Pluim. You gave me a lot of freedom. Maybe sometimes even too much! But I am extremely grateful for that. It gave me the opportunity to fully explore all ideas (good and bad ones) and to spend time on understanding optimisation theory. Also, you are a great listener. Whenever I came to you with some new idea, you listened with apparent interest, thought it over, and usually found the weak point at once.

Marius Staring then. I think I could not have wished a better colleague and room mate. Both working on the topic of image registration, we had an unlimited number of brainstorm sessions. Also with regard to programming (`elastix` and `praxix`) and correcting manuscripts, we had a great cooperation, although you are quite a "komma neuker" sometimes! But I guess I am not the easiest either in that respect. Besides work, we had a lot of fun. I enjoyed the many discussions on work related stuff, religion, politics, and any other theme, often accompanied by coffee or beer. And fortunately, your skin was thick enough to deal with crude jokes. To me, you became a good friend, and I am glad you will be a paranymph.

I am very much indebted to the people from the Department of Radiotherapy, especially Uulke van der Heide. Before our cooperation started, I was working on theoretical solutions, without actually having a problem to solve. This is a suboptimal situation. You provided me with a challenging clinical problem, which was exactly what I needed. Uulke, the word "lazy" does probably not exist in your dictionary. Whenever I needed image data, manual segmentations, or feedback on results, I got them with the speed of light, which is very motivating. Special thanks go also to Ellen Kerkhof, Irene Lips, and Marco van Vulpen, for creating manual segmentations. Furthermore, I would like to thank Gijs Bol, Alexis Kotte, Richard van der Put, and Bas Raaymakers, for useful discussions.

Within the Image Sciences Institute (ISI) there are a lot of people to be acknowledged. I am therefore bound to forget somebody here. I thank the following people for sharing their software, for taking care of the facilities at ISI, for finding bugs in `elastix`, for fruitful discussions, or just for pleasant company: Adriënne, Arnold, Bas, Bom, Bram, Brendan, Casper, Cynthia, Eva, Everine, Ewoud, Frans, Gerard, Gerrit, Hans, Herke-Jan, Ingrid, Ivana, Jacqueline, Jaco, Jan Henry, Jeroen, Joffrey, Kajo, Keelin, Koen, Lisette, Luuk, Maartje, Marjan, Mark, Marloes, Meindert, Michiel, Nelly, Niels, Patrik, Paul, Peter, Rashindra, Renée, Sandra, Sara, Sascha, Silvia, Theo, Thijs, Yulia, Volkan, Wilbert, Wiro. At this point I also would like to thank my new colleagues at the Biomedical Imaging Group Rotterdam for the very nice atmosphere there.

Life outside work becomes pleasant with family and friends. My parents A3 & Monika, brother Martin, Sneek, Bitter Hærvest, W5, Hondsrug, Inge, Marije, Marjolijn, Gala & Krzysztof, just to name a few. My good friend Marno Panis will be a paranymph, thereby representing Sneek. He convinced me to move to Utrecht, which was a good move I think. Good that you convinced me, although you probably did it just in your own interest, to have a place to crash after partying in Utrecht, haha! Finally, MAGDA, you are simply the best.

Stefan Klein
Utrecht, August 3, 2008.

Contents

Chapter 1

Introduction

Image registration is an optimisation problem. Methods for solving this optimisation problem are studied in the present thesis. Thereby, we focus on medical applications of image registration. This chapter introduces the general context, defines the research objectives, and gives an overview of the rest of the thesis.

Image registration is the problem of finding a coordinate transformation that spatially aligns two or more images. It is a common necessity in applications of medical imaging. The images involved can be from different modalities, different time points, and/or different subjects. Extensive surveys on methods for image registration can be found in [9, 26, 39, 43, 47, 48, 99].

This thesis focusses on the registration of pairs of images. One of the images, called the *moving* image, is deformed to fit the other image, the *fixed* image. The quality of alignment is defined by a cost function $\mathcal{C}$, which measures the similarity of the fixed image and the deformed moving image. A high similarity leads to a low cost function and vice versa. An example of a cost function is the mean squared difference of voxel intensities. The coordinate transformation that relates the fixed and moving image is estimated by iteratively minimising the cost function. This means that *image registration is an optimisation problem*. Figure 1.1 illustrates the process. The point at the top of the figure, with high cost function value, corresponds to the unaligned images. The solution of the registration problem is the point at the bottom of the graph, with minimum cost function value. The arrows represent the iterative procedure to find the minimum. In the example of Fig. 1.1 the transfor-

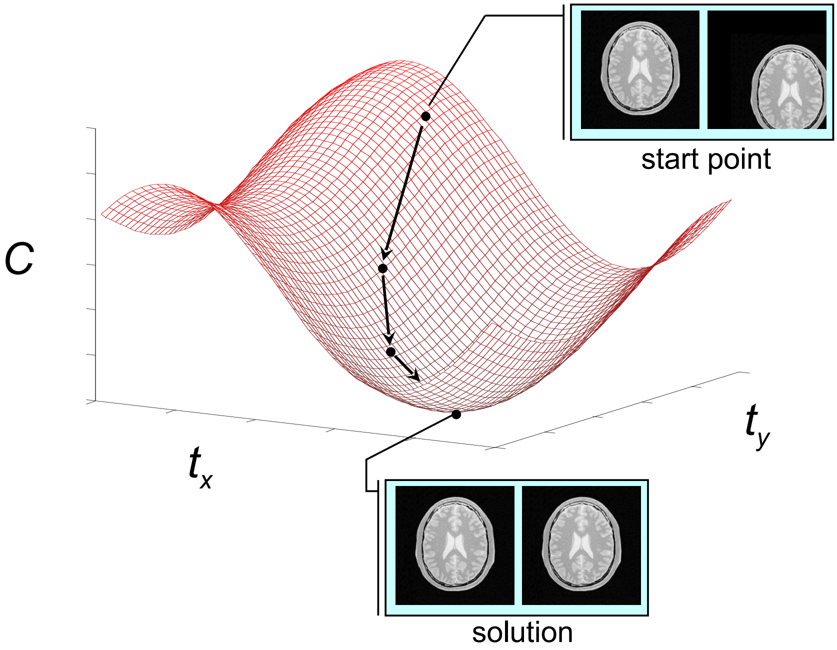
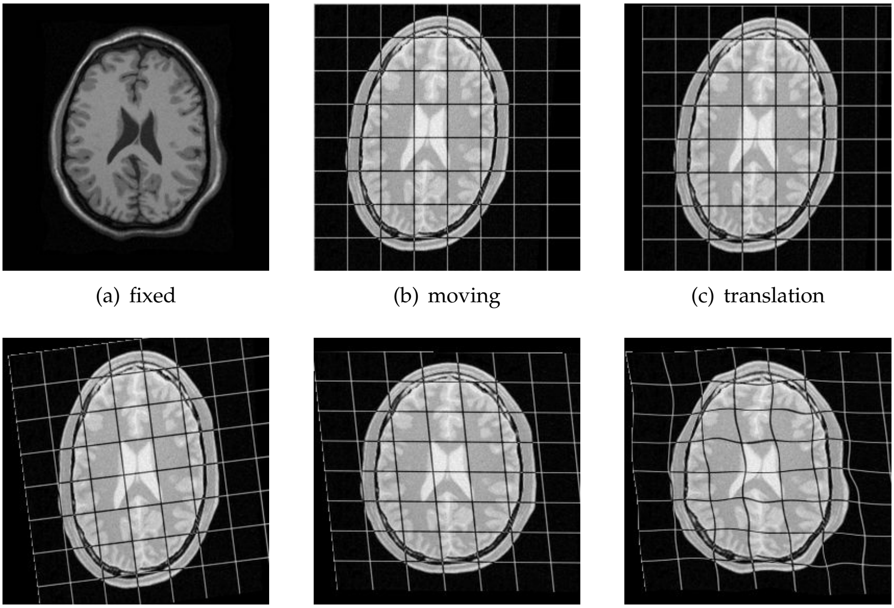


Figure 1.1: *Iterative optimisation. Example for registration with a translation transformation model. The t_x and t_y axes correspond to the translations in x and y direction, respectively. The arrows indicate the steps taken in the direction of the optimum, which is the minimum of the cost function $\mathcal{C}$.*



(a) fixed

(b) moving

(c) translation

(d) rigid

(e) affine

(f) B-spline

Figure 1.2: *Different transformation models. (a) the fixed image, (b) the moving image with a grid overlayed, (c) the deformed moving image with a translation transformation, (d) a rigid transformation, (e) an affine transformation, and (f) a B-spline transformation. The deformed moving image nicely resembles the fixed image using the B-spline transformation. The overlay grids give an indication of the deformations imposed on the moving image.*

mation has only two degrees of freedom: translation in x and y direction. In many medical applications this is not sufficient. Rotations, global scaling, skew, or local deformations are often necessary to align the images. Figure 1.2 shows the effect of increasing the number of degrees of freedom. Only in Fig. 1.2(f) was the transformation model flexible enough to allow for an accurate match. This means that many applications of image registration result in a high-dimensional optimisation problem, with often more than 10000 variables. This, together with the usually nonlinear nature of the cost function, makes image registration a nontrivial problem.

The fact that image registration is mathematically formulated as an optimisation problem has one advantage: there is a vast amount of literature on the topic of optimisation methods. Mathematical optimisation problems have been studied since ages ago. Examples of famous scientists that worked on optimisation are Newton and Lagrange. Nowadays, optimisation is still one of the most studied topics in mathematics. Through the years, many algorithms have been proposed, often accompanied by theoretical analysis of the convergence properties and an extensive experimental validation. Well-known examples are gradient descent, conjugate gradient, (quasi-) Newton, Powell, simulated annealing, and evolutionary strategies.

As always, the good news comes with a downside. The availability of those many optimisation methods triggers the question of "which algorithm to use for my specific application?". Unfortunately, the literature does not provide a definite answer to this question. General guidelines may be given, based on theory or experience, but no single rule can be extracted. The choice of optimisation method is very much linked to the characteristics of the cost function. A noisy cost function, with many small local minima and maxima, demands another algorithm than a nicely smooth function, for example. This makes the choice of optimisation method application dependent.

Besides the challenge of choosing the proper algorithm, there is a second issue that requires attention: parameter setting. Most optimisation methods rely on some user-defined parameters, such as the maximum number of iterations, the "step size" in each iteration, and the convergence criterion. Incorrect settings of these parameters may lead to an excessive computation time, low precision, or even divergence of the solution. Theoretical analysis of the method's convergence properties can provide some guidance, but frequently the mathematical conditions for convergence are hard to translate to the problem at hand.

Regarding our image registration problem we are thus facing two important research questions:

- Which optimisation method to use?
- How can we estimate reasonable values for the user-defined parameters (if any) of the optimisation method?

These questions are the main topic of this thesis. Special attention is paid to optimisation methods that employ stochastic subsampling techniques to reduce the computation time.

The outline of the thesis is as follows. Chapter 2 presents `elastix`, a software package for medical image registration, which was used for all registration experiments described in this thesis. The chapter also gives a general overview of image registration. In Chapter 3, an evaluation of existing optimisation methods is presented. The chapter focusses on mono- and multimodal nonrigid image registration, using a B-spline transformation model and *mutual information* as a cost function. The experimental results indicate that a stochastic gradient descent algorithm (Robbins-Monro) performs best. Chapters 4 and 5 present extensions of the Robbins-Monro algorithm, aimed at simplifying parameter selection and further acceleration. The method in Chapter 4 is designed for monomodal image registration, both rigid and nonrigid, using the mean squared difference as a cost function. The method described in Chapter 5 is aimed at a wider range of applications: rigid, affine, and nonrigid transformation models are considered, and four different cost functions are tested. After these three rather technical chapters on optimisation methods, Chapter 6 discusses an application of medical imaging where efficient nonrigid image registration is an important prerequisite: atlas-based segmentation of the prostate. The chapter is more clinically oriented than the other chapters and has a strong focus on validation. A part of this chapter describes a new optimisation technique for a localised version of mutual information. The thesis ends with a summary and discussion in Chapter 7.

Chapter 2

elastix

S. Klein[†], M. Staring[†], K. Murphy, M.A. Viergever, and J.P.W. Pluim. elastix: a toolbox for intensity-based medical image registration. *Submitted*.

Medical image registration is an important task in medical image processing. It refers to the process of aligning data sets, possibly from different modalities (e.g., magnetic resonance (MR) and computed tomography (CT)), different time points (e.g., follow-up scans), and/or different subjects (in case of population studies). A large number of methods for image registration are described in the literature. Unfortunately, there is not one method that works for all applications. This chapter presents elastix*: a publicly available computer program for intensity-based medical image registration. The software consists of a collection of algorithms that are commonly used to solve medical image registration problems. The modular design of* elastix *allows the user to quickly configure, test, and compare different registration methods for a specific application. The command-line interface enables automated processing of large numbers of data sets, by means of scripting. The usage of* elastix *for comparing different registration methods is illustrated with three example experiments, in which individual components of the registration method are varied.*

[†] In alphabetical order. Both authors contributed equally.

2.1 Introduction

Image registration is a frequently used technique in medical image processing. It is the task of finding the spatial relationship between two or more images. Areas of application include the alignment of data sets from different modalities [46], comparison of follow-up scans to a base-line scan [77], alignment of pre- and postcontrast images [65, 68, 77], updating treatment plans for radiotherapy and surgery [22, 55], atlas-based segmentation [64, Chapter 6 of this thesis], creating models of anatomy [81], and aligning training images for classification [1, 93].

In registration, one image, which is called the *moving image* $I_M(\boldsymbol{x})$, is deformed to fit the other image, the *fixed image* $I_F(\boldsymbol{x})$. In other words, registration is the problem of finding a *coordinate transformation* $\boldsymbol{T}(\boldsymbol{x})$ that makes $I_M(\boldsymbol{T}(\boldsymbol{x}))$ spatially aligned with $I_F(\boldsymbol{x})$. The quality of alignment is defined by a cost function $\mathcal{C}(\boldsymbol{T}; I_F, I_M)$. The optimal coordinate transformation is estimated by minimising the cost function with respect to $\boldsymbol{T}$, usually by means of an iterative optimisation method embedded in a hierarchical (multiresolution) scheme. Extensive reviews on the subject of image registration are given in [9, 26, 39, 43, 48].

Application of an image registration method requires many choices to be made, such as the the optimisation method, the multiresolution strategy, the method of image interpolation to evaluate $I_M(\boldsymbol{T}(\boldsymbol{x}))$, the coordinate transformation model, and the definition of the cost function. Several possibilities for the optimisation method are discussed in [42] and in Chapter 3 of this thesis. An overview of multiresolution strategies is given in [39]. Various image interpolation methods are compared in [59]. The degrees of freedom of the coordinate transformation $\boldsymbol{T}$ determine the types of deformations that can be recovered. Whereas in many applications it may be sufficient to consider only rigid transformations (global translations and rotations) [92, 95], frequently a more flexible transformation model is needed, allowing for local deformations [5, 21, 46, 68]. For the cost function $\mathcal{C}$ many options have been proposed in the literature. Commonly used intensity-based cost functions are the mean squared difference (MSD) [37, 84], normalised correlation (NC) [56, 79], mutual information (MI) [41, 47, 85, 92], and normalised mutual information (NMI) [64, 65, 80]. Sometimes, a regularisation term is added to the cost function, in order to penalise undesired deformations [21, 65, 77]. In medical image processing research it is often necessary to compare several options for each of the registration components. Given the large number of choices, this can be a tedious procedure.

This chapter presents `elastix`: a software package for medical image registration. The `elastix` software has a modular design, including several optimisation methods, multiresolution schemes, interpolators, transformation models, and cost functions. This allows the user to quickly compare different registration methods, in order to select a satisfactory configuration for a specific application. `elastix` has a command-line interface, which enables automated processing of large numbers of data sets, by means of scripting. The software is built upon a widely used open source library for medical image processing, the Insight Toolkit (ITK) [28].

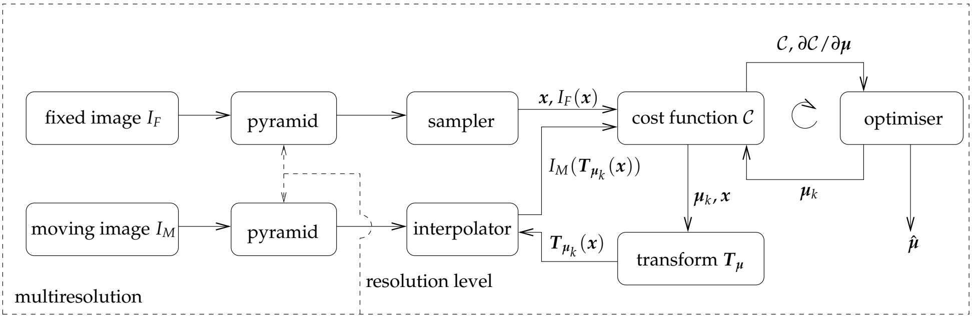


Figure 2.1: *The basic registration components. The scheme is an extended version of the scheme introduced in [28].*

In Sec. 2.2, the general registration framework of `elastix` is discussed, key features of the software are presented, and an overview of the available registration components is given. In Sec. 2.3, three examples of applications that can be handled with the software are given. In these experiments, the influence of three important registration components is demonstrated.

2.2 Image registration with `elastix`

2.2.1 Registration framework

Mathematically, the registration problem is formulated as an optimisation problem in which the cost function $\mathcal{C}$ is minimised with respect to $\boldsymbol{T}$. The `elastix` software is based on a parametric approach, meaning that the number of possible transformations is limited by introducing a parametrisation of the transformation. The optimisation problem reads:

$$\hat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu}} \mathcal{C}(\boldsymbol{T}_{\boldsymbol{\mu}}; I_F, I_M), \tag{2.1}$$

where the subscript $\boldsymbol{\mu}$ indicates that the transform has been parameterised. The vector $\boldsymbol{\mu}$ contains the transformation parameters. The reader is referred to [21, 48] for an overview on nonparametric methods. The minimisation problem (2.1) is solved with an iterative optimisation method, usually in a multiresolution setting. A schematic overview of the basic registration components and their relations is given in Fig. 2.1, which is a slightly extended version of the scheme introduced in [28].

2.2.2 Software characteristics

The `elastix` software is structured according to the block scheme of Fig. 2.1. For each component (transform, cost function, etc.) several choices are available. The user can configure a registration algorithm by specifying the names of the desired components in a parameter text file. Additional settings that some components may require can also be entered in this parameter file. Fixed and moving image file names are supplied as command-line arguments, so that multiple image pairs can be registered using the same parameter settings.

All output of the registration, such as the deformed moving image $I_M(\boldsymbol{T}_{\hat{\mu}}(\boldsymbol{x}))$ and intermediate progress information, is saved to disk. It is often necessary to apply the resulting transformation $\boldsymbol{T}_{\hat{\mu}}$ to data sets other than the moving image. For example, in atlas-based segmentation methods [64, Chapter 6 of this thesis] the transformation is applied to a segmentation (label image) of the moving image. To that end, `elastix` outputs a text file that describes the transformation $\boldsymbol{T}_{\hat{\mu}}$. This text file can subsequently be passed to an accompanying program, called `transformix`, together with the image to be deformed. This program can also be used to evaluate the transformation at user-defined points, or to generate the deformation field.

A large part of the `elastix` code is based on the ITK [28]. The use of the ITK implies that the low-level functionality (image classes, memory allocation etc.) is thoroughly tested. Naturally, all image formats supported by the ITK are supported by `elastix` as well. The source code can be compiled on multiple operating systems, using various compilers, and supports both 32 and 64 bit systems. Executables and source code are publicly available from the website `http://elastix.isi.uu.nl`, under the BSD license. A manual for `elastix` and an example of usage can also be downloaded. The manual includes an example parameter file, describes in detail the various options that can be specified, and provides recommendations for image registration.

2.2.3 Registration components

In the following subsections, more information is given about each component of the block scheme in Fig. 2.1.

2.2.3.1 Cost function

The cost function $\mathcal{C}$ measures the similarity between the fixed image and the deformed moving image. An example is the MSD:

$$\mathrm{MSD}(\boldsymbol{T}_{\mu}; I_F, I_M) = \frac{1}{N} \sum_{\boldsymbol{x} \in \Omega_F} \left(I_F(\boldsymbol{x}) - I_M(\boldsymbol{T}_{\mu}(\boldsymbol{x})) \right)^2, \tag{2.2}$$

where Ω_F denotes the fixed image domain, and N the number of voxels $\boldsymbol{x}$ sampled from the fixed image domain. The sampler, which is responsible for selecting the samples $\boldsymbol{x}$, is discussed in more detail in Sec. 2.2.3.4.

The following metrics are currently supported by `elastix`: MSD, NC, MI, NMI, and the κ-statistic. MSD is only suited for two images with equal intensities, i.e. for images from the same modality. NC is less strict, it assumes an affine relation between the intensity values of the fixed and moving image. MI and NMI assume only a statistical relation between the intensities of the images. They are therefore suited not only for monomodal, but also for multimodal image pairs. The κ-statistic can be used for registering binary images. It measures the overlap of objects in the images. For each of the metrics above, a localised version can be constructed, as proposed in Chapter 6, by selecting the appropriate sampler. This is described in Sec. 2.2.3.4.

Parameters such as the number of bins of the joint histogram, needed for MI and NMI, can be set in the aforementioned parameter file.

When a nonrigid transformation model is used, a regularisation term that penalises undesired deformations can be added to the cost function. An example is the incompressibility constraint described by [65], which penalises compression and expansion of structures. Other examples of regularisation terms are the bending energy of a thin plate [68] and the rigidity penalty term [77]. `elastix` supports these constraints, but currently in combination with MI only.

2.2.3.2 Transformation

The parametrisation of the coordinate transformation $\boldsymbol{T}_{\mu}$ determines the degrees of freedom of the deformation. An example is the affine transformation model, which allows for translation, rotation, scaling and skew of the images:

$$\boldsymbol{T}_{\mu}(\boldsymbol{x}) = A\boldsymbol{x} + \boldsymbol{t}, \tag{2.3}$$

where A is a matrix and $\boldsymbol{t}$ represents the translation vector. The parameter vector $\boldsymbol{\mu}$ is formed by the matrix elements a_{ij} and the translation vector. In 2D, this gives a vector of length 6: $\boldsymbol{\mu} = (a_{11}, a_{12}, a_{21}, a_{22}, t_x, t_y)^T$. In 3D, $\boldsymbol{\mu}$ consists of 9 matrix elements and 3 translations.

The following transformation models are currently supported by `elastix`: translation, rigid (translation and rotation), similarity (rigid plus isotropic scaling), affine, and nonrigid. In the literature, several nonrigid transformation models have been proposed [3, 63, 68], each having its own advantages and disadvantages. In `elastix` a B-spline representation [68] has been implemented. The transformation is modelled as a weighted sum of B-spline basis functions, placed on a uniform control point grid. The B-spline basis functions have local support [86], which is beneficial for fast computation. The flexibility of the deformation is defined by the resolution of the control point grid, which has to be supplied by the user via the parameter file. Section 2.3.1 demonstrates the effect of using different transforms for an example application.

Frequently, nonrigid registration must be preceded by a rigid or affine registration, in order to achieve a rough initial alignment. `elastix` supports the concatenation of any sequence of transforms. The user may also supply an initial transformation, determined in advance, for example by manually clicking corresponding points.

2.2.3.3 Optimisation

To solve (2.1), an iterative optimisation procedure is employed. In every iteration k, the current transformation parameters $\boldsymbol{\mu}_k$ are updated by taking a step in the search direction $\boldsymbol{d}_k$:

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - a_k \boldsymbol{d}_k, \tag{2.4}$$

with a_k a scalar that determines the step size. A wide range of optimisation methods can be formulated in this way, each having different definitions of a_k and $\boldsymbol{d}_k$ (see Chapter 3). A common choice for the search direction is the derivative of the cost function $\partial \mathcal{C} / \partial \boldsymbol{\mu}$ evaluated at the current position $\boldsymbol{\mu}_k$. In this case, Eq. (2.4) boils down to a gradient descent method.

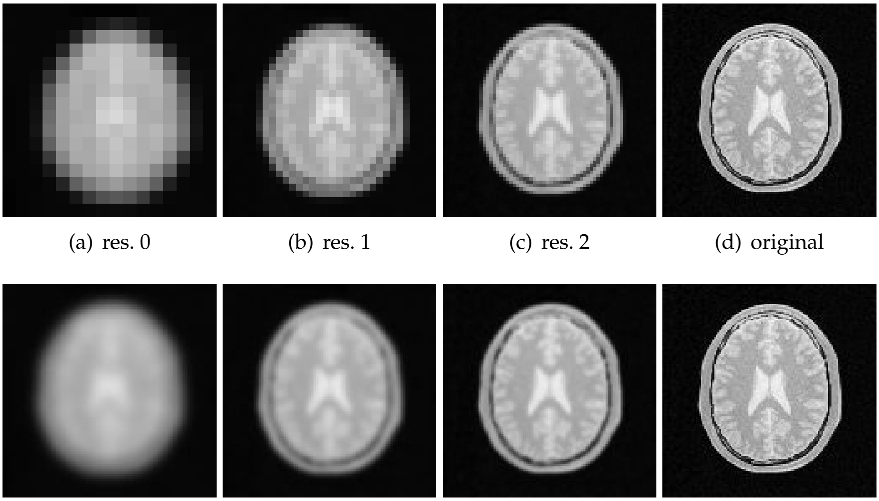
`elastix` includes all optimisation methods described in this thesis: gradient descent, quasi-Newton, nonlinear conjugate gradient (several variants), evolution strategy, and a number of stochastic gradient descent methods (Kiefer-Wolfowitz, Robbins-Monro, simultaneous perturbation, adaptive stochastic gradient descent, and preconditioned stochastic gradient descent). Also an exhaustive search routine is included, which is mainly useful for examining the cost function, as demonstrated in Sec. 2.3.2. The experimental results described in the rest of this thesis indicate that a stochastic gradient descent method (Robbins-Monro, or one of its enhanced versions described in Chapters 4 and 5) is a good choice for many applications. It reduces the computation time per iteration by using only a small subset of the fixed image's voxels for computing the cost function derivative. In each iteration, new samples must be selected randomly. This can be realised in `elastix` by selecting an appropriate sampler, which is explained in the next subsection.

2.2.3.4 Sampling strategies

To compute the cost function $\mathcal{C}$ (and its derivative $\partial \mathcal{C} / \partial \boldsymbol{\mu}$) a set of samples $\boldsymbol{x} \in \Omega_F$ needs to be selected, as in (2.2). The sampler component in Fig. 2.1 is responsible for this. The most straightforward strategy is to use all voxels from the fixed image, which has as an obvious downside that it is time-consuming for large images. A common approach is to use a subset of voxels, selected on a uniform grid, or sampled randomly. Another strategy is to pick only those points that are located on striking image features, such as edges.

`elastix` currently supports the use of all voxels, a subset of voxels selected on a uniform grid, random sampling of voxels, and random sampling off the voxel grid (at non-voxel locations). Random sampling off the grid has been shown to improve the smoothness of the cost function [40, 82]. In Sec. 2.3.2, we demonstrate this effect by comparing several sampling schemes on an example application. For all sampling strategies discussed above, the user may optionally supply a mask image, indicating regions of interest. In this way, one can force the sampler to pick only points near edges in the image, for example.

With random sampling, the `elastix` user can enforce the selection of new samples in every iteration k of the optimisation process. In this way, the stochastic optimisation methods described in Chapters 3-5 can be realised. The localised mutual information strategy, presented in Chapter 6, can be implemented by letting the sampler pick points in a small neighbourhood. A new neighbourhood is randomly selected in every iteration of the optimisation procedure.



(a) res. 0 (b) res. 1 (c) res. 2 (d) original

(e) res. 0 (f) res. 1 (g) res. 2 (h) original

Figure 2.2: *Two multiresolution strategies using a Gaussian pyramid ($\sigma = 8.0, 4.0, 2.0$ voxels). The top row shows multiresolution with downsampling, the bottom row without. Note that in the top row the number of voxels in each dimension is halved every resolution, but the voxel size is doubled, so physically the images are of the same size.*

2.2.3.5 Interpolation

For computation of the cost function, the value $I_M(\boldsymbol{T}_{\mu}(\boldsymbol{x}))$ is evaluated at non-voxel positions, for which intensity interpolation is needed. Several methods for interpolation exist, varying in quality and speed, including nearest neighbour, linear and N-th order B-spline interpolation [83, 86]. `elastix` supports all interpolators mentioned above.

2.2.3.6 Hierarchical strategies

Hierarchical (multiresolution) strategies are an important aspect of image registration. For an extensive overview, the reader is referred to [39]. `elastix` implements several hierarchical strategies.

The pyramid components in the block scheme of Fig. 2.1 represent the multiresolution schemes for the *image data*. Two types of image pyramids are available in `elastix`: Gaussian pyramids with and without downsampling. Figure 2.2 illustrates the difference.

The second important multiresolution strategy, not apparent from Fig. 2.1, is the gradual increase of *transformation model* complexity. During nonrigid registration, a hierarchical effect can be realised by starting with a coarse B-spline control point resolution and gradually refining the grid in subsequent resolutions, thereby introducing the capability to recover more fine-scale deformations.

More generally, any parameter setting can be subjected to a hierarchical strategy in elastix. For example, the number of joint histogram bins that is used for computing MI and NMI could be gradually increased, as was suggested in [85].

In Sec. 2.3.3 several multiresolution strategies are compared for the nonrigid registration of CT chest scans.

2.3 Experiments and results

In this section, some applications of elastix are described, to illustrate its convenience for configuring, testing, and comparing different registration methods. Three key components of registration were studied: the transformation model in Sec. 2.3.1, the sampling technique in Sec. 2.3.2, and the multiresolution strategy in Sec. 2.3.3. The experiments demonstrate the impact these components can have on the registration results, and therefore stress the importance of a proper configuration for the application at hand.

2.3.1 Transformation models

The effect of the type of transformation was investigated by comparing the registration performance of several transformation models.

To this end, a set of 50 clinical MR scans of the prostate was used, all originating from different patients. The scans were made by the Department of Radiotherapy of the University Medical Center Utrecht, as part of prostate cancer treatment planning. They were acquired on a Philips 3T scanner (Gyroscan NT Intera, Philips Medical Systems, Best, The Netherlands) using a balanced steady-state free precession sequence with fat suppression. The scans had a dimension of $512 \times 512 \times 90$ voxels of size $0.49 \times 0.49 \times 1.0$ mm.

Fifty interpatient registrations were performed by registering each MR scan with its predecessor in the 50-scan series. Interpatient registration of these scans is needed for atlas-based segmentation of the prostate (see Chapter 6). The registration problem is challenging, since the anatomical variability between subjects is large. Also, the data suffer from several artefacts, as will be shown in Chapter 6. For our experiments we used the same settings as in Chapter 6, with localised MI as a cost function (see Sec. 2.2.3.4), and a four-level Gaussian image pyramid with downsampling. The following transformation models were compared: translation, rigid, affine, and B-spline with different control point spacings: 64, 32, 16, 8, and 4 mm. The result of the registration with translations only was used as an initialisation for all other registrations. For the B-spline registrations, the control point grid was subjected to a multiresolution scheme: registration starts with a coarse control point resolution; with smoother versions of the images, the control point resolution is increased accordingly.

For all images a manual segmentation of the prostate was available, made by an experienced radiation oncologist. After registration with elastix, the transformation $\boldsymbol{T}_{\hat{\mu}}$ was applied to the prostate segmentation of the moving image, using

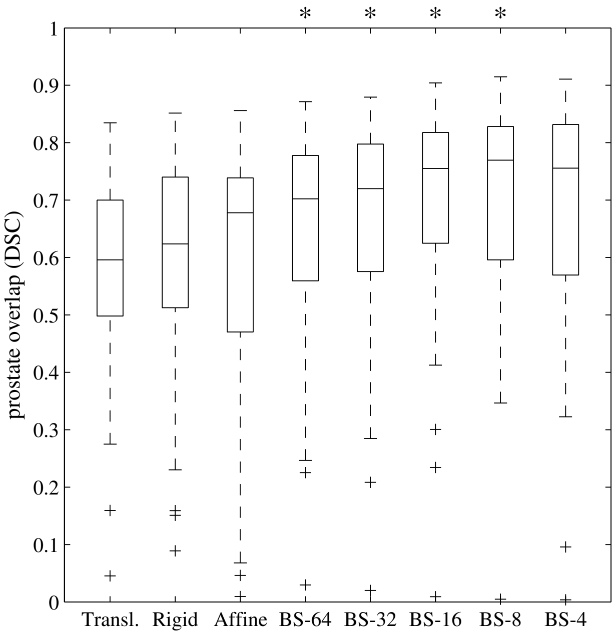


Figure 2.3: *The effect of the transformation model on the accuracy of registration, measured by the prostate overlap. The abbreviation BS-$\langle sp \rangle$ refers to a B-spline transformation with control point spacing sp, in mm.*

`transformix`. The overlap with the segmentation of the fixed image was computed, using the Dice similarity coefficient (DSC) [18]:

$$\text{DSC}(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|}, \tag{2.5}$$

where X and Y represent the two segmentations, and $| \cdot |$ denotes the number of voxels within the segmentation. A DSC of 1 indicates perfect registration. A value of 0 means that the prostates had no overlap at all after registration.

The results are presented in Fig. 2.3. For each transformation model, the DSC values of the 50 MR scans were summarised by a box-and-whiskers plot. A paired, two-sided Wilcoxon test was used to assess the median differences between adjacent columns. A star on top of a column indicates a significant difference ($p < 0.05$) with respect to the previous column. The graph clearly shows that a nonrigid registration was essential in this application. The best results were obtained using a B-spline control point spacing of 8 mm. With this setting, the computation time was around 15 minutes per registration on a single processor Pentium 2.8 GHz personal computer.

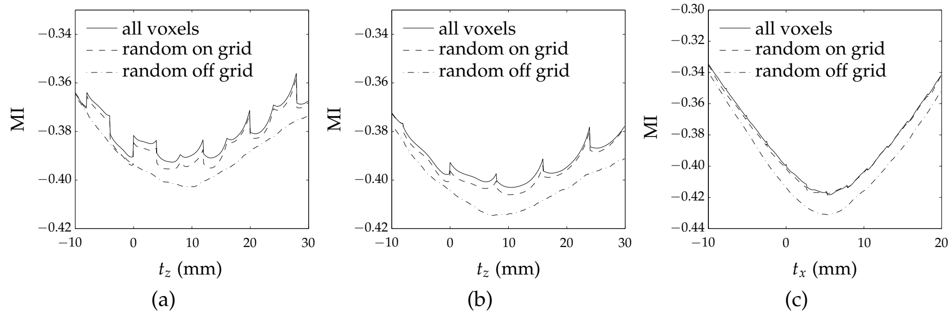


Figure 2.4: *The effect of different sampling strategies on the smoothness of the cost function; a) translation in the z direction, b) translation in the z direction after downsampling the MR image in the z direction, c) translation in the x direction.*

2.3.2 Sampling strategies

The grid effect is a well-known issue in image registration. It refers to the problem that the cost function contains irregularities at locations representing grid-aligning transformations, which can impede the registration process. It has often been studied in the context of interpolation artefacts [59]. In this section it is demonstrated that the sampling mechanism can solve this issue, by taking samples off the voxel grid, as suggested in [40, 82].

Brain images were taken from the "Retrospective Image Registration Evaluation" project [95]. We investigated the registration of a T1-weighted MR image (moving image) to a PET image (both of patient 001). The PET image had a dimension of $128 \times 128 \times 15$ voxels of size $2.59 \times 2.59 \times 8.0$ mm. The MR image had a dimension of $256 \times 256 \times 26$ voxels of size $1.25 \times 1.25 \times 4.0$ mm.

The cost function (MI) was analysed using an exhaustive search in a single translation direction, with a step size of 0.1 mm. Linear interpolation was used to compute $I_M(\boldsymbol{T}_{\boldsymbol{\mu}}(\boldsymbol{x}))$. Different sampling strategies were employed for computing the cost function: all voxels, random sampling at the voxel grid, and random sampling off the voxel grid.

In Fig. 2.4(a) the cost function $-\mathrm{MI}(\boldsymbol{T}_{\boldsymbol{\mu}}; I_F, I_M)$ is plotted as a function of the translation t_z, the direction with the largest voxel spacing. The two samplers that take samples on the voxel grid have a very irregular cost function. The irregularities show a pattern, related to the voxel sizes of the images in the z direction (8 mm for PET, 4 mm for MR). Every 8 mm a slice of the PET image maps outside the MR image. This causes the large discontinuities at $t_z = 12$ mm and 20 mm for example. Every 4 mm, the cost function exhibits a small local maximum, caused by the aligning voxel grids of the images. The random sampler that takes samples off the grid clearly leads to a much smoother cost function.

The experiment was repeated after downsampling the MR image by a factor of 2 in the z direction. The voxel size of the MR image thus became equal to that of the PET image (8 mm). Figure 2.4(b) shows the cost function as a function of t_z. The irregularities follow a single pattern in this graph, with a peak at every 8 mm.

Figure 2.4(c) shows the result for translation in the x direction (obtained using the original non-downsampled MR image). The portion of voxels of the PET image that move simultaneously outside the MR image domain is smaller than in the z direction. Consequently, the grid effect is reduced. The cost function appears much smoother, also for the two samplers that take samples on the voxel grid, although small irregularities remain visible at multiples of the voxel spacing. This example shows that, in practice, it may not always be strictly necessary to sample voxels off the grid.

2.3.3 Multiresolution strategies

The influence of the choice of multiresolution strategy is examined in this section. CT chest scans of 26 patients were taken from a lung cancer screening trial [88]. Each patient had a baseline and a follow-up scan, acquired 3-9 months apart. The scans were obtained at full inspiration and without contrast injection on a 16-detector-row scanner (Mx8000 IDT or Brilliance 16P, Philips Medical Systems, Best, The Netherlands). Images were of size 512×512 in-plane, with the voxel size ranging from 0.55×0.55 mm to 0.8×0.8 mm. The number of slices varied from 383 to 529, with slice thickness 1 mm and slice spacing 0.7 mm. The images were downsampled by a factor of two in each dimension before registration, in order to decrease computational load.

For each patient the baseline and follow-up scans were registered using a non-rigid B-spline transformation. An affine registration was used for initialisation. The registration was performed with a Gaussian image pyramid (without downsampling) using $R \in \{1, \ldots, 8\}$ levels. Two experiments were performed for each value of R. Firstly, the resolution of the B-spline control point grid was kept at a constant value of 12 mm (isotropic) in all resolutions. Secondly, the grid was refined after each resolution, such that at the final resolution the control points were spaced 12 mm apart again. This yields 16 experiments on 26 image pairs, resulting in a total of 416 registrations. For the cost function MI was used. The Robbins-Monro stochastic optimisation method was applied, using 1000 iterations per resolution level. The image sampler was configured to select 2000 samples randomly in each iteration.

One hundred corresponding points in each baseline and follow-up scan were established by two independent observers using a semi-automatic algorithm [51]. The transformation $\boldsymbol{T}_{\hat{\mu}}$ was applied to the annotated points in the fixed image using `transformix`. To evaluate the registration accuracy, the mean distance between the resulting locations and the reference standard of the observer annotations was computed.

Figure 2.5 shows box-and-whisker plots of the mean distance to the annotations of one of the observers. The interobserver variability is shown in the left-most column. The first group R1-R8 displays the results without grid refinement. The second group shows the results with grid refinement. When no grid refine-

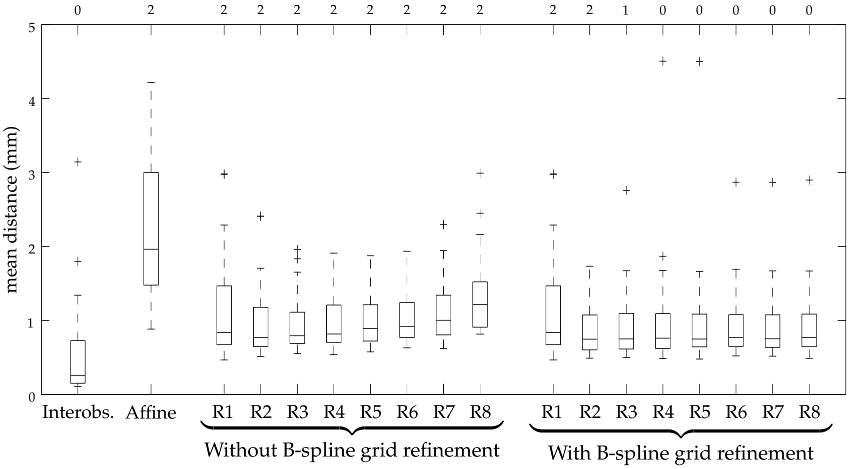


Figure 2.5: *The effect of the multiresolution strategy on registration accuracy, expressed as the mean distance between corresponding points. 'R1' - 'R8' refer to the number of image resolution levels that was used. The numbers on top of the graph refer to the number of outliers with mean distance larger than 5 mm.*

ment was used, the registration quality improved until $R = 3$, but deteriorated for $R > 3$. Apparently, the dense B-spline grid yielded too much freedom on the heavily smoothed images. With grid refinement the results kept improving with increasing R up to $R = 6$ (note the decreasing number of outliers above 5 mm). In practice, when considering the computation time, three or four resolutions with grid refinement seems to be a reasonable choice. With these settings the runtime was about 10 minutes on an AMD Opteron running at 2.4 GHz.

2.4 Conclusion

A software package, `elastix`, for medical image registration has been presented. Rather than implementing a single registration method, `elastix` is a collection of parametric intensity-based registration methods. Thanks to the modular design, the user can easily construct a registration algorithm, tailored to a specific application. Configuration of the registration method can be accomplished by writing a few lines in a parameter text file, without having to write any programming code. `elastix` has a command-line interface, which simplifies batch-processing of large numbers of data sets. Registration of large three-dimensional images can be done efficiently, thanks to the use of stochastic subsampling techniques.

The usage of `elastix` has been illustrated with three experiments. In the first experiment, eight transformation models were compared for the interpatient registration of 50 MR prostate scans. In the second experiment, we reproduced a result from the literature, showing that the so-called grid effect can be reduced by sampling the fixed image off the voxel grid. The third experiment demonstrated the importance of choosing a suitable hierarchical (multiresolution) strategy, by registering 26 chest CT image pairs with 16 different multiresolution configurations. These three investigations are just a few examples of the many possible comparative studies that one can perform with `elastix`.

The software has been used in several research projects, including [31, 32, 76, 77, 89, 93]. Both the executables and the source code are publicly available. The source code provides the users with the exact construction of the available algorithms, and allows them to enhance the functionality of `elastix` by adding their own algorithms. These features, in combination with the modular design, make `elastix` a useful tool for research on medical image registration.

Chapter 3

Evaluation of Optimisation Methods

S. Klein, M. Staring, J.P.W. Pluim. Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines. *IEEE Transactions on Image Processing*, 16(12):2879–2890, 2007.

A popular technique for nonrigid registration of medical images is based on the maximisation of their mutual information, in combination with a deformation field parameterised by cubic B-splines. The coordinate mapping that relates the two images is found using an iterative optimisation procedure. This chapter compares the performance of eight optimisation methods: gradient descent (with two different step size selection algorithms), quasi-Newton, nonlinear conjugate gradient, Kiefer-Wolfowitz, simultaneous perturbation, Robbins-Monro, and evolution strategy. Special attention is paid to computation time reduction by using fewer voxels to calculate the cost function and its derivatives. The optimisation methods are tested on manually deformed CT images of the heart, on follow-up CT chest scans, and on MR scans of the prostate acquired using a BFFE, T1, and T2 protocol. Registration accuracy is assessed by computing the overlap of segmented edges. Precision and convergence properties are studied by comparing deformation fields. The results show that the Robbins-Monro method is the best choice in most applications. With this approach, the computation time per iteration can be lowered approximately 500 times without affecting the rate of convergence by using a small subset of the image, randomly selected in every iteration, to compute the derivative of the mutual information. From the other methods the quasi-Newton and the nonlinear conjugate gradient method achieve a slightly higher precision, at the price of larger computation times.

3.1 Introduction

Nonrigid registration is an important technique in medical image processing. However, in general it requires a large computation time, which is a big disadvantage for many clinical applications. Comprehensive studies, such as lung cancer screenings, with many high resolution 3D images, ask for faster registration algorithms [73]. Other applications, such as brain shift estimation based on intraoperatively acquired ultrasound [55], require almost real-time registration. Also in external radiotherapy there is a need for fast registration methods. Movements of organs may cause discrepancies between the expected radiation dose distribution and the actually received dose. Fast nonrigid registration would allow for on-line updating of the treatment plan [20].

The aim of registration is to find a deformation field $\boldsymbol{u}$ that spatially relates two images, such that the deformed 'moving' image $I_M(\boldsymbol{x} + \boldsymbol{u}(\boldsymbol{x}))$ matches the 'fixed' image $I_F(\boldsymbol{x})$ at every position $\boldsymbol{x}$. In this work we focus on a widely used nonrigid registration technique, based on maximisation of the *mutual information* similarity measure, in combination with a deformation field parameterised by *cubic B-splines* [46, 68]. The approach can be formulated as a minimisation problem:

$$\hat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu}} \mathcal{C}\left(\boldsymbol{\mu}; I_F, I_M\right), \tag{3.1}$$

where the cost function $\mathcal{C}$ equals the negated mutual information similarity metric, and $\boldsymbol{\mu}$ represents the parameter vector containing the B-spline coefficients that define the deformation field $\boldsymbol{u}$. The cost function $\mathcal{C}$ may have multiple local minima. Which local minimum is selected as the solution $\hat{\boldsymbol{\mu}}$ depends on the optimisation algorithm and on the initial alignment of the images. A regularisation term $\mathcal{R}$ can be added to the cost function, to penalise undesirable deformations, and, consequently, to reduce the number of local minima:

$$\mathcal{C}(\boldsymbol{\mu}; I_F, I_M) = -\mathrm{MI}(\boldsymbol{\mu}; I_F, I_M) + \omega\mathcal{R}(\boldsymbol{\mu}). \tag{3.2}$$

In this equation ω serves as a weighting factor for the regularisation term. Well-known examples for $\mathcal{R}$ are the curvature term [21], the elastic energy [10], and the volume preserving penalty term [65]. With nonparametric registration techniques, which do not employ a parametric model of the deformation, a proper regularisation term is essential to ensure smoothness (differentiability) of the deformation field [21]. In the parametric approach that we focus on, the regularisation term may be superfluous, since the cubic B-spline basis functions are inherently smooth. However, additional regularisation may be needed in order to, for instance, avoid singularities ('folding effects') in the deformation field.

To determine the optimal set of parameters $\hat{\boldsymbol{\mu}}$ an iterative optimisation strategy is employed:

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k + a_k \boldsymbol{d}_k, \quad k = 0, 1, 2, \ldots, \tag{3.3}$$

with $\boldsymbol{d}_k$ the 'search direction' at iteration k, and a_k a scalar gain factor controlling the step size along the search direction. The search directions and gain factors are chosen such that the sequence $\{\boldsymbol{\mu}_k\}$ converges to a local minimum of the cost function $\mathcal{C}$. Many optimisation methods can be found in the literature [6, 35, 36, 53], differing in the way a_k and $\boldsymbol{d}_k$ are computed. In contrast to the field of *rigid* registration [42], no extensive comparison of optimisation procedures has been done for *nonrigid* image registration problems.

In this work, several optimisation methods are compared with respect to speed, accuracy, precision, and robustness. The following methods are included in the study: gradient descent [53], quasi-Newton [17], nonlinear conjugate gradient [16], Kiefer-Wolfowitz [30], simultaneous perturbation [74], Robbins-Monro [62], and evolution strategy [23]. The first three are deterministic gradient-based algorithms. They have in common that the expression for the search direction $\boldsymbol{d}_k$ is based on $\partial\mathcal{C}/\partial\boldsymbol{\mu}$, the derivative of the cost function with respect to the parameters, and they assume that $\partial\mathcal{C}/\partial\boldsymbol{\mu}$ can be computed exactly. The second three methods are stochastic gradient-based algorithms. They also derive their search directions from $\partial\mathcal{C}/\partial\boldsymbol{\mu}$, but only need *stochastic approximations* of the derivative, potentially faster to compute than the exact derivative. The last method, evolution strategy, is not based on $\partial\mathcal{C}/\partial\boldsymbol{\mu}$, but it can be classified as stochastic, since its choice of search directions depends on a random process. Section 3.3 explains the optimisation methods under scrutiny.

Special attention is paid to the effect of using only a small, randomly selected set of image samples in each iteration, instead of the full image. This is an easy way to decrease the computation time per iteration, but it may deteriorate the rate of convergence. The stochastic nature of such an approximation technique makes it unsuitable for the deterministic optimisation methods, because they expect exact derivatives. However, stochastic optimisation methods may be able to deal with it. The technique has been proposed for rigid registration problems [92], but its effect on *nonrigid* registration has not been evaluated in the literature. Section 3.4 discusses the topic more extensively.

The experiments and results are described in Sec. 3.5. The optimisation methods are tested on manually deformed CT images of the heart, on follow-up CT scans of the chest, and on MR scans of the prostate acquired with three different protocols. Conclusions are given in Sec. 3.6.

3.2 Nonrigid registration method

This section describes the various components of the nonrigid registration method. The design of the algorithm is largely based on the papers by Rueckert et al. [68], Mattes et al. [46], and Thévenaz and Unser [85].

The registration method uses cubic B-splines to parameterise the deformation field, and mutual information as the similarity measure. Several implementations for the computation of mutual information can be found in the literature [41, 79, 85, 92]. The approach described by Thévenaz and Unser [85] is used here. The mutual information is defined as follows:

$$\mathrm{MI}(\boldsymbol{\mu}; I_F, I_M) = \sum_{m \in L_M} \sum_{f \in L_F} p(f, m; \boldsymbol{\mu}) \log_2 \left(\frac{p(f, m; \boldsymbol{\mu})}{p_F(f) p_M(m; \boldsymbol{\mu})} \right), \tag{3.4}$$

where L_F and L_M are sets of regularly spaced intensity bin centres, p is the discrete joint probability, and p_F and p_M are the marginal discrete probabilities of the fixed and moving image, obtained by summing p over m and f, respectively. The joint probabilities are estimated using B-spline Parzen windows:

$$\begin{aligned} p(f, m; \boldsymbol{\mu}) = \frac{1}{|I_F|} \sum_{\boldsymbol{x}_i \in I_F} & w_F(f/\sigma_F - I_F(\boldsymbol{x}_i)/\sigma_F) \\ & \times w_M(m/\sigma_M - I_M(\boldsymbol{x}_i + \boldsymbol{u}_{\boldsymbol{\mu}}(\boldsymbol{x}_i))/\sigma_M), \end{aligned} \tag{3.5}$$

where $\boldsymbol{x}_i$ denotes the spatial coordinates of voxel i in the fixed image volume I_F, $\boldsymbol{u}_{\boldsymbol{\mu}}$ is the B-spline deformation field, and w_F and w_M represent the fixed and moving Parzen windows. For w_M a third-order B-spline is used, which makes it possible to derive an analytic expression for $\partial \mathrm{MI}/\partial \boldsymbol{\mu}$, see [46, 85]. For w_F a zeroth-order B-spline can be used [46]. The scaling constants σ_F and σ_M must equal the intensity bin widths defined by L_F and L_M. These follow directly from the grey-value ranges of I_F and I_M and the user-specified number of histogram bins $|L_F|$ and $|L_M|$.

A number of experiments described in this chapter have been performed with and without the regularisation term $\mathcal{R}$. A regularisation term is used that penalises second order derivatives of the deformation field:

$$\mathcal{R} = \frac{1}{|I_F|} \sum_{\boldsymbol{x}_i \in I_F} \sum_{p,q,r} \left(\frac{\partial^2 u_p}{\partial x_q \partial x_r}(\boldsymbol{x}_i) \right)^2. \tag{3.6}$$

Equivalent combinations of q and r that occur twice are counted once.

To guide the optimisation towards the desired local minimum of the cost function, multiresolution strategies are often employed. For extensive overviews on this subject we refer to [38, 39]. In our experiments the commonly used Gaussian image pyramid was used for the image data. The complexity of the deformation model is defined by the B-spline control point resolution. We let it follow the image resolution: when the image resolution is doubled, the control point resolution is doubled as well. The number of resolution levels and the final B-spline control point spacing are problem specific.

3.3 Optimisation methods

All optimisation algorithms studied in this chapter can be written in the form of (3.3). The methods differ in the way they compute the gain factors a_k and search directions $\boldsymbol{d}_k$.

Many strategies exist for determining the gain a_k. It can, for example, simply be set to a constant, or defined by a decaying function of k. Another possibility is the use of a *line search*, which, in each iteration, tries to minimise the cost function $\mathcal{C}$ along the search direction $\boldsymbol{d}_k$:

$$a_k = \arg\min_a \mathcal{C}\left(\boldsymbol{\mu}_k + a\boldsymbol{d}_k\right). \tag{3.7}$$

The disadvantage of such an *exact* line search is that many additional evaluations of the cost function and/or its derivative are required. Therefore, an *inexact* line search is more often used. Instead of solving (3.7) exactly, an inexact line search finds a gain factor a_k that gives a sufficient reduction of $\mathcal{C}$.

In all but one of the investigated optimisation methods, the expression for $\boldsymbol{d}_k$ is based on the derivative of the cost function, $\partial\mathcal{C}/\partial\boldsymbol{\mu}$, henceforth referred to as $\boldsymbol{g}$. As mentioned in Sec. 3.2, an analytic expression for the derivative of the mutual information is available. Some optimisation methods require exact evaluation of this expression. Other methods are satisfied with an approximation.

3.3.1 Gradient descent (GDD and GDL)

The gradient descent method [53] takes steps in the direction of the negative gradient of the cost function:

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - a_k\boldsymbol{g}(\boldsymbol{\mu}_k), \tag{3.8}$$

where $\boldsymbol{g}(\boldsymbol{\mu}_k)$ is the derivative of the cost function evaluated at the current position $\boldsymbol{\mu}_k$.

In this work we study two variants of the gradient descent method. In the first variant, called GDD, the gain factor a_k is defined as a decaying function of k: $a_k = a/(k+A)^\alpha$, with user-defined constants $a > 0$, $A \geq 1$, and $0 \leq \alpha \leq 1$. This choice makes the gradient descent method more comparable to the *stochastic* gradient descent algorithms, see Sec. 3.3.4, where the specific form of this expression is justified. In the second variant, called GDL, the gain factor is determined by an inexact line search routine, called 'Moré-Thuente'. This choice makes the gradient descent method more comparable to the quasi-Newton and nonlinear conjugate gradient methods, which are described in Secs. 3.3.2 and 3.3.3. Further details about the Moré-Thuente algorithm are given in those sections.

In order to give an indication of the rate of convergence of gradient descent methods, it is possible to derive theoretical bounds on the distance to the solution at iteration k, $||\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}}||$. Provided that the sequence $\{\boldsymbol{\mu}_k\}$ converges to a local nonsingular minimum $\hat{\boldsymbol{\mu}}$ of $\mathcal{C}$, it can be proven [6] that there exist a $K \geq 0$ and $\rho > 0$ such that the following expression holds:

$$\frac{||\boldsymbol{\mu}_{k+1} - \hat{\boldsymbol{\mu}}||}{||\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}}||} \leq \rho, \quad \text{for all } k \geq K. \tag{3.9}$$

This means that the method has a linear rate of convergence. If $\rho \geq 1$, the term 'sublinear convergence' is used [29].

3.3.2 Quasi-Newton (QN)

Quasi-Newton methods [17, 53] are inspired by the well-known Newton-Raphson algorithm, which is given by:

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - [H(\boldsymbol{\mu}_k)]^{-1}\boldsymbol{g}(\boldsymbol{\mu}_k), \tag{3.10}$$

where $H(\boldsymbol{\mu}_k)$ is the Hessian matrix of the cost function, evaluated at $\boldsymbol{\mu}_k$. The use of such second order information gives the algorithm better theoretical convergence properties than the gradient descent. The computation of the Hessian matrix and its inverse is computationally expensive, especially in high-dimensional optimisation problems such as nonrigid registration. Quasi-Newton methods tackle this problem by using an approximation to the inverse of the Hessian: $L_k \approx [H(\boldsymbol{\mu}_k)]^{-1}$. The approximation is updated in every iteration k. Second order derivatives of the cost function are *not* needed for this update; only the already computed first order derivatives are used. Direct approximation of the inverse of the Hessian avoids the need for a matrix inversion. Quasi-Newton methods are typically implemented in combination with an inexact line search routine, determining a gain factor a_k that ensures sufficient progress towards the solution. This results in the following QN algorithm:

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - a_k L_k \boldsymbol{g}(\boldsymbol{\mu}_k). \tag{3.11}$$

Given certain conditions many quasi-Newton methods can be shown to be *superlinearly* convergent [17]:

$$\lim_{k\to\infty} \frac{||\boldsymbol{\mu}_{k+1} - \hat{\boldsymbol{\mu}}||}{||\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}}||} \to 0. \tag{3.12}$$

Many ways to construct the series $\{L_k\}$ are proposed in the literature [17, 53], most notably Symmetric-Rank-1 (SR1), Davidon-Fletcher-Powell (DFP), and Broyden-Fletcher-Goldfarb-Shanno (BFGS). Numerical experiments indicate that BFGS is very efficient in many applications [53]. It uses the following update rule for L_k:

$$L_{k+1} = \left(I - \frac{\boldsymbol{s}\boldsymbol{y}'}{\boldsymbol{s}'\boldsymbol{y}}\right) L_k \left(I - \frac{\boldsymbol{y}\boldsymbol{s}'}{\boldsymbol{s}'\boldsymbol{y}}\right) + \frac{\boldsymbol{s}\boldsymbol{s}'}{\boldsymbol{s}'\boldsymbol{y}}, \tag{3.13}$$

where the accent denotes the transpose operation, I is the identity matrix, $\boldsymbol{s} = \boldsymbol{\mu}_{k+1} - \boldsymbol{\mu}_k$, and $\boldsymbol{y} = \boldsymbol{g}_{k+1} - \boldsymbol{g}_k$. In our study we use a popular variant of the BFGS method, the Limited memory BFGS (LBFGS) [52], which eliminates the need for storing the matrix L_k in memory.

Following the implementation described in [52] we use the inexact line search routine described by Moré and Thuente [49]. It determines a_k such that the so-called *strong Wolfe conditions* are satisfied:

$$\mathcal{C}(\boldsymbol{\mu}_{k+1}) \leq \mathcal{C}(\boldsymbol{\mu}_k) + c_1 a_k \boldsymbol{d}_k' \boldsymbol{g}(\boldsymbol{\mu}_k), \tag{3.14}$$

$$|\boldsymbol{d}_k' \boldsymbol{g}(\boldsymbol{\mu}_{k+1})| \leq c_2 |\boldsymbol{d}_k' \boldsymbol{g}(\boldsymbol{\mu}_k)|, \tag{3.15}$$

with user-defined scalars c_1 and c_2 satisfying $0 < c_1 < c_2 < 1$. Recall that $\boldsymbol{d}_k$ represents the search direction of the optimisation algorithm, see (3.3), which equals

$-L_k \boldsymbol{g}(\boldsymbol{\mu}_k)$ in the case of quasi-Newton methods. The first Wolfe condition (3.14) demands a sufficient decrease of the cost function value. The second Wolfe condition (3.15) enforces reasonable progress towards a stationary point of the cost function, where the derivative vanishes. For optimisation problems where the computational cost of evaluating the gradient $\boldsymbol{g}_k$ is high compared to the cost of computing L_k, the values $c_1 = 10^{-4}$ and $c_2 = 0.9$ are suggested in [60]. To realise superlinear convergence it is important to always try a gain factor $a_k = 1$ first [53]. If this step size does not satisfy the strong Wolfe conditions, the iterative Moré-Thuente line search procedure is started to find a suitable gain. If no gain factor satisfying the strong Wolfe conditions can be found, the optimisation is assumed to have converged.

3.3.3 Nonlinear conjugate gradient (NCG)

The development of conjugate gradient methods started with the *linear* conjugate gradient method [25]. This routine was designed for solving a system of linear equations, which is equivalent to the minimisation of a quadratic cost function. The *nonlinear* conjugate gradient method is an extension suitable for minimising general nonlinear functions [16, 53]. The NCG algorithm follows the general iterative scheme (3.3). The search direction $\boldsymbol{d}_k$ is defined as a linear combination of the gradient $\boldsymbol{g}(\boldsymbol{\mu}_k)$ and the previous search direction $\boldsymbol{d}_{k-1}$:

$$\boldsymbol{d}_k = -\boldsymbol{g}(\boldsymbol{\mu}_k) + \beta_k \boldsymbol{d}_{k-1}. \tag{3.16}$$

Several expressions for the scalar β_k have been proposed in the literature [16], including:

$$\text{Dai-Yuan: } \beta_k^{\mathrm{DY}} = \frac{\boldsymbol{g}_k' \boldsymbol{g}_k}{\boldsymbol{d}_{k-1}'(\boldsymbol{g}_k - \boldsymbol{g}_{k-1})}, \tag{3.17}$$

$$\text{Hestenes-Stiefel: } \beta_k^{\mathrm{HS}} = \frac{\boldsymbol{g}_k'(\boldsymbol{g}_k - \boldsymbol{g}_{k-1})}{\boldsymbol{d}_{k-1}'(\boldsymbol{g}_k - \boldsymbol{g}_{k-1})}, \tag{3.18}$$

where the notation $\boldsymbol{g}_k = \boldsymbol{g}(\boldsymbol{\mu}_k)$ is introduced for clarity. The choice of β_k has a large influence on the global convergence properties. For an extensive review on this topic we refer to [16]. In our study we use a hybrid version, proposed in [15] and shown to be very efficient compared to other methods:

$$\beta_k = \max\left(0, \min\left(\beta_k^{\mathrm{HS}}, \beta_k^{\mathrm{DY}}\right)\right). \tag{3.19}$$

Depending on the line search technique used, various theoretical bounds on the rate of convergence have been derived in the literature. Most results are obtained assuming an exact line search. In practice, an exact line search is seldom feasible, since it would require too many cost function evaluations. In [50] it is shown that, with a more practical inexact line search routine, a superlinear rate of convergence can be achieved. For our comparative study, we choose the same inexact line search routine as used with the QN method, i.e. the Moré-Thuente algorithm. Whereas the unit gain has to be tried first for QN, there is no such rule for NCG. A reasonable approach is to try $a_k = a_{k-1}$ as a first guess. This choice appears to satisfy the strong Wolfe conditions often, and thus inhibits the number of line search iterations needed. For the GDL method, see Sec. 3.3.1, the same approach is used.

3.3.4 Stochastic gradient descent

The stochastic gradient descent method [36] follows the same scheme as the *deterministic* gradient descent, see (3.8), with the distinction that the derivative of the cost function, $\boldsymbol{g}(\boldsymbol{\mu}_k)$, is replaced by an approximation $\widetilde{\boldsymbol{g}}_k$, resulting in the following scheme:

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - a_k \widetilde{\boldsymbol{g}}_k. \tag{3.20}$$

Convergence to the solution $\hat{\boldsymbol{\mu}}$ can only be guaranteed [35] if the bias of the approximation error goes to zero:

$$\mathrm{E}(\widetilde{\boldsymbol{g}}_k) \rightarrow \boldsymbol{g}(\boldsymbol{\mu}_k), \quad \text{as } k \rightarrow \infty, \tag{3.21}$$

where $\mathrm{E}(\cdot)$ denotes expectation. A stochastic gradient descent method is often applied when computation of the exact derivative is very costly. Using an approximation of the exact derivative could decrease the computation time per iteration, but may have negative effects on the speed of convergence.

Three variants of the stochastic gradient method are investigated in this chapter: Kiefer-Wolfowitz, simultaneous perturbation, and Robbins-Monro.

Kiefer-Wolfowitz (KW) This method, originally proposed in [30], is based on a finite difference approximation of the derivative, given by:

$$[\widetilde{\boldsymbol{g}}_k]_i = \frac{\mathcal{C}\left(\boldsymbol{\mu}_k + c_k \boldsymbol{e}_i\right) - \mathcal{C}\left(\boldsymbol{\mu}_k - c_k \boldsymbol{e}_i\right)}{2c_k}, \tag{3.22}$$

where $[\widetilde{\boldsymbol{g}}_k]_i$ represents the ith element of $\widetilde{\boldsymbol{g}}_k$, c_k is a small scalar, and $\boldsymbol{e}_i$ is the unit vector consisting of only zeros, except for the ith element, which equals one. The Kiefer-Wolfowitz method assumes that only approximations of the cost function values are available:

$$\begin{aligned} \widetilde{\mathcal{C}}_{ki}^{+} &= \mathcal{C}\left(\boldsymbol{\mu}_k + c_k \boldsymbol{e}_i\right) + \varepsilon_{ki}^{+} \quad \text{and} \\ \widetilde{\mathcal{C}}_{ki}^{-} &= \mathcal{C}\left(\boldsymbol{\mu}_k - c_k \boldsymbol{e}_i\right) + \varepsilon_{ki}^{-}, \end{aligned} \tag{3.23}$$

where ε_{ki}^{+} and ε_{ki}^{-} represent the approximation errors. Substituting this in (3.22) yields the KW algorithm:

$$[\widetilde{\boldsymbol{g}}_k]_i = \frac{\widetilde{\mathcal{C}}_{ki}^{+} - \widetilde{\mathcal{C}}_{ki}^{-}}{2c_k}. \tag{3.24}$$

The derivative approximation is twofold. Besides the approximation error introduced by the finite difference scheme, an external source of error is taken into account, which is expressed by the ε-terms in (3.23). For c_k the following expression is commonly used:

$$c_k = c/(k+1)^{\gamma}, \tag{3.25}$$

where $c > 0$ and $0 \leq \gamma \leq 1$ are user-defined constants. Note that, for an N-dimensional optimisation problem, the KW procedure requires $2N$ evaluations of the cost function for each iteration k. However, in our application, the computational costs can be reduced by exploiting the compact support of the cubic B-splines that model the deformation field.

Simultaneous perturbation (SP) The simultaneous perturbation method, first described by Spall [74], also bases its derivative estimate on approximate evaluations of the cost function. However, whereas the KW algorithm requires $2N$ cost function evaluations per iteration, the SP method uses only 2 evaluations, independent of N:

$$\begin{aligned} \widetilde{\mathcal{C}}_k^{+} &= \mathcal{C}\left(\boldsymbol{\mu}_k + c_k \boldsymbol{\Delta}_k\right) + \varepsilon_k^{+} \quad \text{and} \\ \widetilde{\mathcal{C}}_k^{-} &= \mathcal{C}\left(\boldsymbol{\mu}_k - c_k \boldsymbol{\Delta}_k\right) + \varepsilon_k^{-}. \end{aligned} \tag{3.26}$$

In these expressions $\boldsymbol{\Delta}_k$ denotes the 'random perturbation vector' of which each element is randomly assigned ±1 in each iteration, with equal probability. The approximation errors are represented by the ε-terms. The ith element of the derivative vector $\widetilde{\boldsymbol{g}}_k$ is then computed by:

$$[\widetilde{\boldsymbol{g}}_k]_i = \frac{\widetilde{\mathcal{C}}_k^{+} - \widetilde{\mathcal{C}}_k^{-}}{2c_k\,[\boldsymbol{\Delta}_k]_i}. \tag{3.27}$$

The scalar c_k is defined according to (3.25). The simultaneous perturbation method has been used for rigid registration [11], but its performance has not been compared to other optimisation methods.

Robbins-Monro (RM) Whereas KW and SP construct a derivative estimate based on approximate evaluations of the cost function, the Robbins-Monro algorithm [62] does not specify how the derivative is computed. It assumes that an approximation of the derivative of the cost function is available:

$$\widetilde{\boldsymbol{g}}_k = \boldsymbol{g}(\boldsymbol{\mu}_k) + \boldsymbol{\varepsilon}_k. \tag{3.28}$$

In fact, this makes KW and SP special cases of RM. Note that, if the $\boldsymbol{\varepsilon}_k$-term is zero in every iteration, the method equals the deterministic gradient descent procedure, described in Sec. 3.3.1.

In Sec. 3.4 a method to approximate the mutual information and its derivative is discussed, which is used in conjunction with KW, SP, and RM in our experiments.

The approximated gradient $\widetilde{\boldsymbol{g}}_k$ does not necessarily vanish close to the solution $\hat{\boldsymbol{\mu}}$, in contrast to the exact derivative that satisfies $\boldsymbol{g}(\hat{\boldsymbol{\mu}}) = 0$. Thus, convergence of $\{\boldsymbol{\mu}_k\}$ must be forced by ensuring $a_k \to 0$ as $k \to \infty$. In most theoretical work on stochastic approximation algorithms, a_k is defined as a decaying function of k: $a_k = a/(k+1)^\alpha$, where $a > 0$ and $0 \leq \alpha \leq 1$ are user-defined constants. In practice the following modified expression is often used [75]:

$$a_k = a/(k+A)^\alpha, \tag{3.29}$$

with $A \geq 1$. This will be used in our experiments. The same gain sequence is used by the GDD method (Sec. 3.3.1). Theoretically optimal values for α are derived in [35] and [74]. For SP specifically, practical guidance for choosing a, A, and α is provided in [75]. For α the lowest theoretically admissible value of 0.602 is recommended. For A a value of approximately 10% of the user-defined maximum number of iterations is suggested, or less. The choice of the overall gain, a, depends on the expected ranges of $\boldsymbol{\mu}$ and $\boldsymbol{g}$ and is thus problem specific.

Due to the stochastic nature of the algorithms, theoretical bounds on the rate of convergence can not be given in the same form as in the previous sections, like in (3.9). Instead, the theoretical convergence properties are given in terms of the 'asymptotic normality' of $(\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}})$:

$$(k+1)^\beta \left(\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}}\right) \sim \mathcal{N}\left(\boldsymbol{m}, \Sigma\right), \quad \text{as } k \to \infty, \tag{3.30}$$

where $\mathcal{N}(\boldsymbol{m}, \Sigma)$ denotes a multivariate normal distribution with mean $\boldsymbol{m}$ and covariance matrix Σ. Intuitively, the higher β, the better the rate of convergence. More details can be found in [19, 35, 69, 74].

3.3.5 Evolution strategy (ES)

Evolution strategies are based on the principle of natural selection. Many variants of the basic idea have been described in the literature. For an extensive review we refer to [8]. The covariance matrix adaptation (CMA) ES [23] is generally considered to be the current state-of-the-art ES algorithm [4], and is therefore included in this study.

Each iteration of the CMA-ES algorithm consists of three phases: offspring generation, selection, and recombination. In the first phase a set of λ trial search directions is generated from a normal distribution $\mathcal{N}$:

$$\boldsymbol{d}_k^{(\ell)} \sim \mathcal{N}(\boldsymbol{0}, C_k), \quad \text{for } \ell = 1, 2, \ldots, \lambda. \tag{3.31}$$

The population size λ is a user-defined parameter. The covariance matrix C_k favours search directions that were successful in previous iterations. For each trial search direction the cost function value $\mathcal{C}(\boldsymbol{\mu}_k + a_k \boldsymbol{d}_k^{(\ell)})$ is evaluated. The scalar a_k again serves the role of a gain factor that controls the step size. The selection phase consists of selecting the $P \leq \lambda$ trial directions that yield the lowest cost function values. The pth best trial direction out of all λ trial directions is denoted by $\boldsymbol{d}_k^{(p;\lambda)}$. In the recombination phase a weighted average of the P selected trial directions is computed:

$$\boldsymbol{d}_k = \sum_{p=1}^{P} w_p \boldsymbol{d}_k^{(p;\lambda)}. \tag{3.32}$$

The weight factors w_p should satisfy $\sum_p w_p = 1$ and $w_p \geq w_{p+1} > 0$. The new position $\boldsymbol{\mu}_{k+1}$ is determined using (3.3).

After each iteration, a_k and C_k are automatically updated, based on the previous search direction $\boldsymbol{d}_{k-1}$ and the selected trial search directions $\boldsymbol{d}_{k-1}^{(p;\lambda)}$. Basically, the gain factor a_k is increased when the preceding search directions are similar, and decreased when the preceding steps tend to cancel each other out. The reader is referred to [23] for the exact adaptation mechanisms of a_k and C_k. The initial step size a_0 is a user-defined parameter. For C_0 the identity matrix is used. Reference [23] also contains expressions for the weights w_p, and gives recommendations for λ and P: $\lambda = 4 + \lfloor 3 \ln N \rfloor$ and $P = \lfloor \lambda/2 \rfloor$, with N the dimension of the parameter vector $\boldsymbol{\mu}$.

Theoretical results on the convergence properties of CMA-ES are not available. For ES algorithms in general some results can be found in [4, 67, 96], for example. Experimental results with synthetic cost functions [2, 7] indicate that approximate (noisy) cost function evaluations can be dealt with to some degree.

3.4 Approximation by subsampling

In this section, we describe two techniques to approximate the mutual information and its derivatives. The approximation techniques are based on subsampling.

In our implementation, the computation times of both the mutual information (t_C) and its derivative (t_g) are linearly dependent on the number of voxels $|I_F|$ in the fixed image. The computation time of the derivative also depends linearly on the number of B-spline coefficients N:

$$t_C \sim p|I_F| + q, \tag{3.33}$$

$$t_g \sim r|I_F| + sN, \tag{3.34}$$

where p, q, r, and s are positive constants. For most nonrigid registration problems $p|I_F|$ tends to be much larger than q, and $r|I_F|$ much larger than sN. In these cases, we can lower the computation time significantly by not using all the voxels, but only a small subset of voxels.

The stochastic optimisation algorithms (KW, SP, RM, and ES) take into account that only approximations of the cost function are available. By using a *new, randomly selected* subset of voxels *in every iteration* of the optimisation process, a bias in the approximation error is avoided. This technique, which we call 'stochastic subsampling', has been proposed before for rigid registration problems [92], but its effect on *nonrigid* registration has not been evaluated in the literature. In our experiments, we test the stochastic algorithms with and without stochastic subsampling.

The number of samples used in each iteration is denoted by a number behind the optimisation method's name. For example, KW-2048 refers to the Kiefer-Wolfowitz method using 2048 voxels. Voxels are allowed to be selected more than once. If all voxels are used to compute the search direction (no subsampling), the postfix '-all' is used.

A possible subsampling strategy for the *deterministic* methods (GDD, GDL, QN, and NCG) is to select *a single* subset of voxels in the fixed image and use these samples throughout the registration process [37, 46]. A disadvantage of this 'deterministic subsampling' technique is that convergence to the correct solution cannot be guaranteed, because the approximation error is biased. However, for completeness we include this technique in our experiments. The deterministic subsampling technique is implemented by selecting voxels on a regular grid using identical downsampling factors for each image dimension. The downsampling factor is added as a number behind the optimisation method's name, for example QN-2. A downsampling factor of 1 corresponds to using the full image.

3.5 Experiments and results

To compare the deformation fields $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ resulting from two different optimisation methods, we define the *average displacement distance*:

$$D(\boldsymbol{u}_1, \boldsymbol{u}_2) = \frac{1}{|I_F|} \sum_{x_i \in I_F} \|\boldsymbol{u}_1(\boldsymbol{x}_i) - \boldsymbol{u}_2(\boldsymbol{x}_i)\| \,. \tag{3.35}$$

When the true solution of a registration problem is known (in case of a manually imposed deformation for example), this measure can also be used to compare the results to the ground truth. The proposed (Euclidian) distance measure is appropriate as long as the deformations are reasonably small. For a discussion on distance metrics for deformation fields, we refer to [44].

To compare the registration results in terms of accuracy, we calculate the overlap of segmented structures after their alignment. The overlap of two corresponding volumes V_1 and V_2 is defined as:

$$\text{overlap} = \frac{2|V_1 \cap V_2|}{|V_1| + |V_2|} \cdot 100\%. \tag{3.36}$$

This measure is known as the Dice similarity index [18]. A higher overlap indicates a better alignment of the objects. A value of 1 indicates perfect overlap, a value of 0 means no overlap at all. The sensitivity of the overlap measure depends on the surface-volume ratio of the objects [64]. To increase the sensitivity we compute the morphological gradients of V_1 and V_2 and evaluate the overlap measure on the resulting edge structures. The morphological gradient of an object is defined as its dilation minus its erosion. For the dilation and erosion we use a 3×3×3 kernel.

The computational efficiency of the optimisation method depends on the number of required iterations and the computation time per iteration. The computation

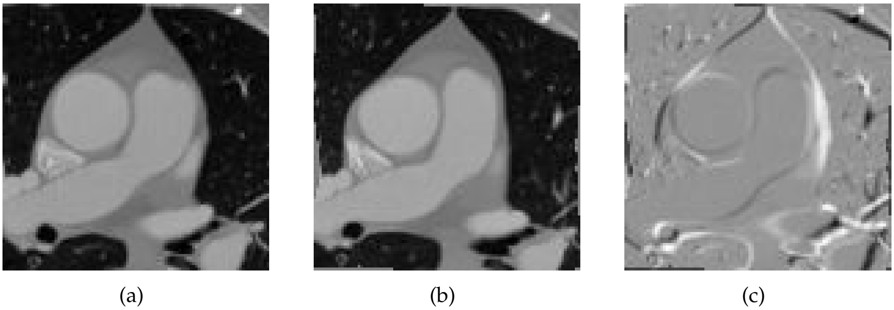

(a) (b) (c)

Figure 3.1: *CT heart data, used in the experiments with known ground truth: (a) an example slice, (b) the same slice after application of the initial deformation field to the image volume, (c) the difference between (a) and (b). Voxels in the deformed volume that map outside the original image were set to 0.*

time per iteration is dominated by the time required for calculating the (approximation of) the mutual information or its derivative. Timing measurements indicated that the term sN in (3.34) can be neglected. Consequently, for GD, QN, NCG, and RM the computation time per iteration equals ηt_g, with η the fraction of the voxels used to compute the derivative, and t_g the time required to compute the derivative using *all* voxels. All timing results in this chapter are reported as a factor times t_g. For example, for 512 iterations of QN-4 with 3D images we report a computation time of $512(1/4)^3 t_g = 8t_g$. Note that the computation times per iteration of KW, SP, and ES are not obviously related to t_g. To express them as a factor times t_g we rely on experimental measurements. For each application we also report the value of t_g in seconds (measured with an AMD Opteron 244, 1.8 GHz), to give an indication of the typical computation times.

3.5.1 Artificial motion

In the first experiment an image I is registered with a deformed version of itself. To avoid interpolation errors, the deformed version of I is not actually generated. Instead, an initial deformation field $\tilde{\boldsymbol{u}}$ is subtracted from the B-spline deformation field $\boldsymbol{u}_\mu$ that is updated during optimisation. The average displacement distance $D(\boldsymbol{u}_\mu, \tilde{\boldsymbol{u}})$ can be used to assess the registration quality.

The registrations were performed on four 3D CT images of the heart. The images originated from chest scans. These were manually cropped to the area of the heart and downsampled by a factor of two in each dimension, resulting in images of 97×97×97 voxels with an isotropic voxel size of 1.4 mm. For each image an initial deformation field $\tilde{\boldsymbol{u}}$ was generated, composed of randomly placed Gaussian blobs with a standard deviation of 14 mm. Each component of $\tilde{\boldsymbol{u}}$ was composed of 300 blobs. The amplitudes of the blobs were uniformly distributed between -3.5 and 3.5 mm. Figure 3.1 shows an example slice, its deformed version, and the difference image visualising the initial misalignment.

The registrations were performed using a 10×10×10 grid of B-spline control points to parameterise the deformation field $\boldsymbol{u}_{\mu}$, yielding $N = 3000$ parameters to be optimised. For the number of histogram bins we used $|L_F| = |L_M| = 32$. No multiresolution schemes were used in this experiment, which makes comparison of the results more straightforward. No regularisation term was used either. The maximum number of iterations was limited to 2000. Three constants must be set for the gain sequence (3.29) employed by the optimisation methods GDD, KW, SP, and RM. For GDD, KW, and RM we used $a = 3200$, $A = 51$, and $\alpha = 0.602$. For SP slightly different parameters had to be used, since the method appeared to be sensitive to the choice of the gain sequence. The following values were used, resulting in a lower gain, especially in the first iterations: $a = 800$, $A = 201$, and $\alpha = 0.602$. Two more parameters need to be specified for KW and SP, see (3.25): $c = 1.0$ and $\gamma = 0.101$. The choices for α and γ are based on the recommendations in [75]. For ES, the initial step size a_0 was set to 1.0, and, following the recommendations in [23], the values $\lambda = 28$ and $P = 14$ were used. The stochastic optimisation methods were tested with and without the stochastic subsampling strategy. Stochastic subsampling was tested using 10^5, 16384, 2048, and 256 voxels. The deterministic methods were tested with the deterministic downsampling strategy, using downsampling factors of 1 (full image), 2, 4, 8, and 16, corresponding to 10^6, 10^5, 15625, 2197, and 343 voxels, respectively.

We present the results for one of the four CT images. The outcome for the other images was similar. In Fig. 3.2 the convergence results are given for all methods, without subsampling. The error measure $D(\boldsymbol{u}_{\mu_k}, \tilde{\boldsymbol{u}})$ is plotted against the number

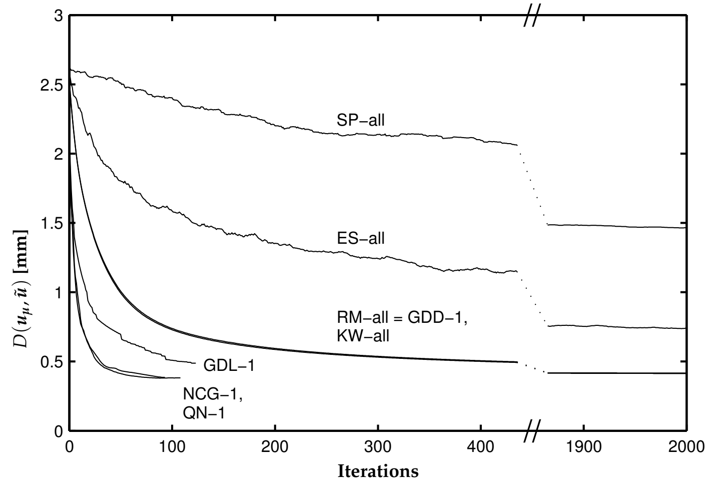
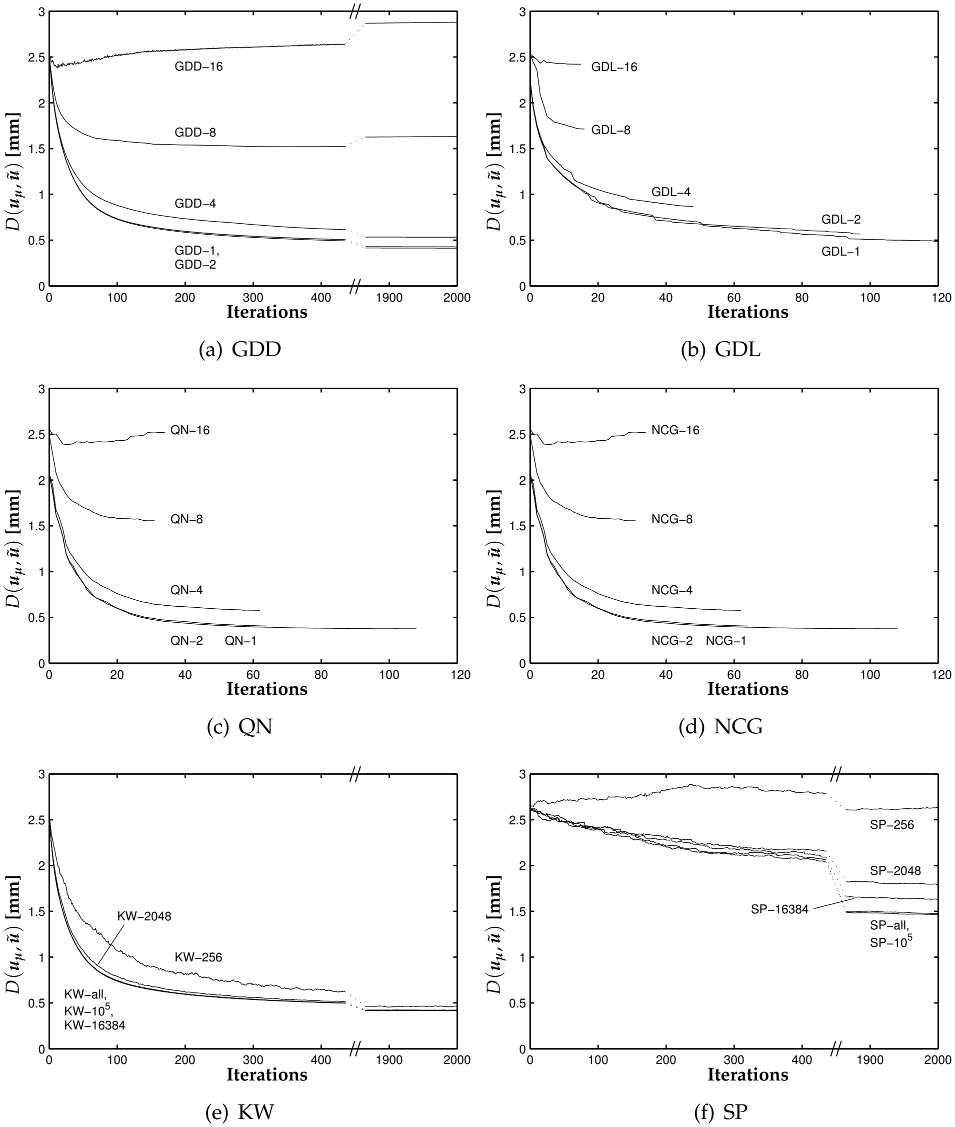


Figure 3.2: *Convergence results for all methods, without subsampling. Note that the horizontal axis contains a gap and that some curves are overlapping.*



(a) GDD

(b) GDL

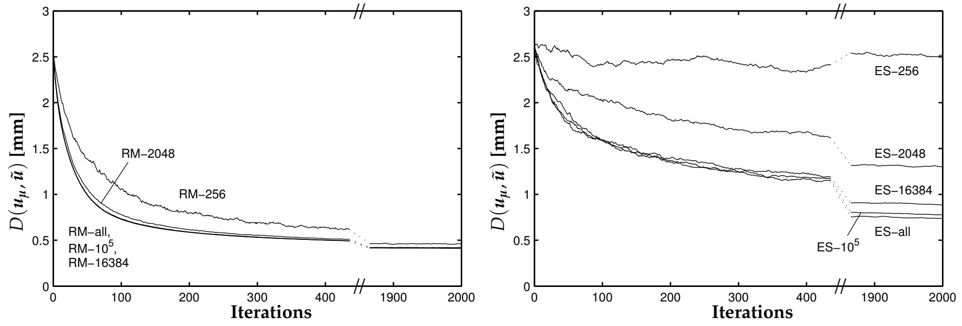
(c) QN

(d) NCG

(e) KW

(f) SP

Figure 3.3: *Convergence results for all methods, with and without subsampling. Note that the horizontal axis contains a gap in some graphs, and does not have the same scale everywhere. Also note that several curves are overlapping.*



(g) RM

(h) ES

Figure 3.3: *Convergence results (continued from previous page).*

of iterations k. The methods GDL, QN, and NCG were terminated before the limit of 2000 iterations was reached, because the strong Wolfe conditions could not be satisfied anymore and convergence was assumed (see Sec. 3.3.2). The graph shows that SP and ES exhibited a substantially lower rate of convergence than the other methods. The methods QN-1 and NCG-1 converged in fewer iterations than the others and achieved a higher precision. The effect of subsampling on the performance of each optimisation method is presented in Fig. 3.3. Figures 3.3(a)-(d) show the effect of deterministic subsampling: downsampling by a factor of 4 or more degraded the registration results of GDD, GDL, QN, and NCG. Figures 3.3(e)-(h) show the results for stochastic subsampling. Interestingly, for RM and KW the convergence properties of using all voxels were retained when going down to only 2048 samples, which is 0.2% of the total image volume. The computation times per iteration of GDD, GDL, QN, NCG, and RM are equal to ηt_g, with η the fraction of voxels used. One t_g was measured to be 20 seconds approximately. For KW, SP, and ES the computation times needed for the cost function evaluations in each iteration were measured to be around $10\eta t_g$, $0.9\eta t_g$ $(= 2\eta t_C)$, and $13\eta t_g$ $(= 28\eta t_C)$, respectively. It follows that the KW method is not competitive, despite its fair rate of convergence. The computation times per iteration of SP and ES do not compensate for their low rates of convergence. Among the stochastic gradient descent methods the RM-2048 procedure clearly performed superior in this experiment. The GDD method with the deterministic downsampling approach is also outperformed by RM. The methods have an equal rate of convergence, but, because of the stochastic subsampling strategy, RM can be used with fewer voxels than GDD with deterministic subsampling. In Fig. 3.4 the average displacement distance is plotted as a function of computation time for the most competitive methods: GDL, QN, and NCG with downsampling factors of 1, 2, and 4, and RM-2048. The result of RM-all is added for reference, to visualise the acceleration realised by stochastic subsampling. The results of GDL-8, QN-8, and NCG-8 are included to show that a downsampling factor higher than four is not feasible for those methods. Note that a

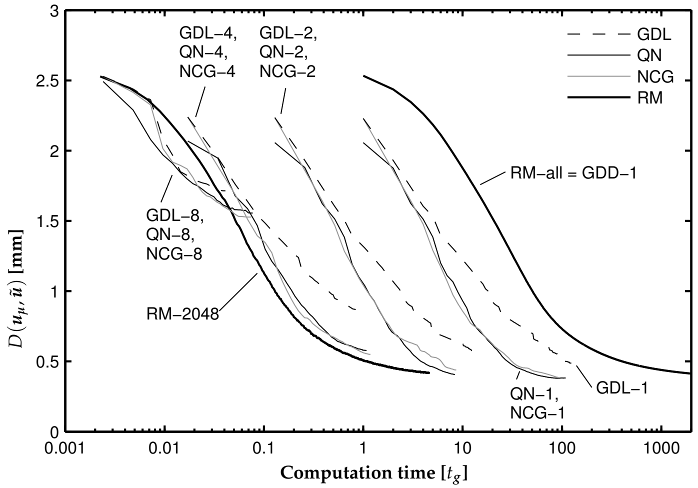


Figure 3.4: *Timing results for GDL, QN, NCG, and RM ($t_g \approx 20$ s.).*

logarithmic scale is used for the horizontal axis. The RM-2048 method is clearly the fastest. The stochastic subsampling strategy yields an acceleration factor of about 500, compared to RM-all. The better rate of convergence of QN and NCG results in an acceleration factor of 10, approximately, compared to RM-all.

The tests were repeated for a more difficult registration problem, constructed by composing the imposed deformation field $\tilde{u}$ of Gaussian blobs with a standard deviation of 7 mm, instead of 14 mm. This smaller standard deviation results in a deformation field that is hard to recover, since the B-spline control point grid used during registration is not dense enough. Each component of $\tilde{u}$ was composed of 1500 blobs. The amplitudes of the blobs were uniformly distributed between -3.5 and 3.5 mm. The timing results for the same CT image as before are shown in Fig. 3.5. Interestingly, the QN and NCG methods could not handle this very ill-defined registration problem. The GDL and RM routines gave reasonable results. As expected, none of the optimisation methods were able to achieve a very large reduction of the initial average displacement error, since the B-spline control point grid was not dense enough. Note that the QN and NCG methods *did* find a set of parameters that decreased the cost function. The Moré-Thuente line search, employed in both QN and NCG to set the gain factor a_k, guarantees that the cost function decreases in every iteration, $\mathcal{C}(\mu_k) < \mathcal{C}(\mu_{k-1})$. Apparently, the lower cost function did not translate into a better accuracy. We have repeated the experiments using the regularisation term $\mathcal{R}$ with $\omega = 500$, see (3.2). This resolved the issue, but did not change the efficiency differences between the methods. The effect of regularisation is studied further in the following sections.

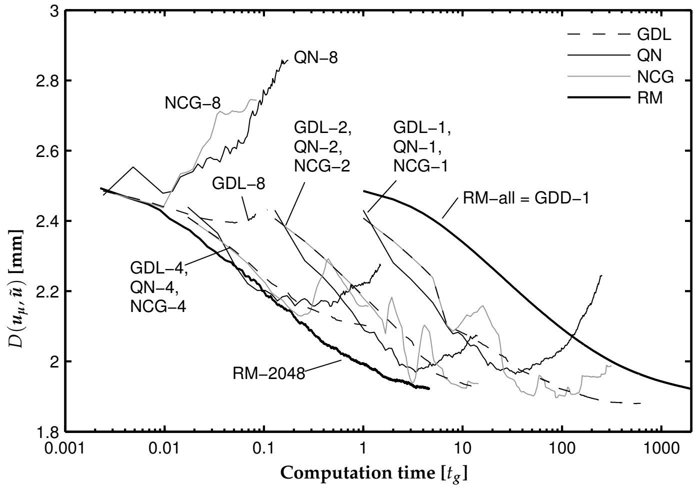


Figure 3.5: *Timing results on a more difficult registration problem ($t_g \approx 20$ s.).*

3.5.2 Motion between follow-up CT chest scans

Computed tomography is a commonly used modality for the diagnosis of lung diseases. To study the evolution of disease in a patient, it is helpful to automatically register follow-up scans. In this section, a number of experiments with follow-up scans of the thorax is described. We limit our attention to the methods that turned out most favourable in the previous section: GDL, QN, NCG, and RM.

The images were acquired with a Philips Mx8000IDT 16-slice CT scanner. The original images, with an in-plane dimension of 512×512 and a number of slices ranging from 400 to 800, were downsampled by a factor of two in each dimension, in order to be able to register the images on a standard PC with one gigabyte of memory. The resulting voxel size was approximately 1.4 mm in all directions. In this study we used data of five patients.

For each patient two scans, taken several months apart, were registered. The nonrigid registration was preceded by a rigid registration with mutual information as the similarity measure. For both rigid and nonrigid registration a four-level multiresolution approach was applied. At each resolution the number of iterations was limited to 1000. At the highest resolution the B-spline control point spacing was set to 22 mm, yielding a grid of about 19×19×19 control points; approximately 20000 parameters to optimise. For the number of histogram bins we used $|L_F| = |L_M| = 32$. The RM method was tested with and without stochastic subsampling. The numbers of voxels used with the stochastic subsampling strategy were 10^5, 16384, 2048, and 256 voxels. The GDL, QN, and NCG methods were tested with downsampling factors of 1, 2, 4, 8, and 16, respectively corresponding to about 10^7, 10^6, 10^5, 20000, and 2500 voxels. For the gain sequence a_k the following parameters were used: $a = 40000$, $A = 51$, and $\alpha = 0.602$.

Experiments were performed both with and without the regularisation term $\mathcal{R}$. For the weighting factor ω a value of 500 was used. Without a regularisation term QN and NCG yielded unrealistic deformation fields at low-contrast regions of the image. The Jacobian of the transformation $\boldsymbol{x} + \boldsymbol{u}(\boldsymbol{x})$ exhibited large negative values, indicating foldings in the deformation field. With a regularisation term the foldings were avoided. The RM procedure did not have this problem. It produced a folding only once, in the vicinity of a fast-growing tumour. The GDL method had similar problems as QN and NCG, but to a lesser extent.

To compare the methods in terms of registration accuracy, we use the overlap measure, applied on the morphological gradients of segmentations of the lungs. The segmentations were made by means of a region-growing method based on the work of Hu et al. [27, 73]. Pulmonary vessels are not included in the lung segmentations, so that the morphological gradient of the segmentation contains the vessel boundaries and the global lung boundaries.

The precision is measured by the average displacement distance D to the solution obtained by QN-1, since that method found the deformation with the lowest cost function value and is thus our best estimate of the true optimum. The precision values are calculated on a region of interest defined by dilation of the lung segmentations with a 7×7×7 structuring element.

The results are located in the left part of Table 3.1. Overlap and precision values were calculated after rigid registration and nonrigid registration using GDL, QN, NCG, and RM, all with regularisation. The results for the five patients are summarised by the average (avg) and standard deviation (sd). The first column, 'time', shows the average required computation time for one registration (number of iterations times computation time per iteration). One t_g was measured to be 220 seconds approximately. The time needed to calculate the derivative of the regularisation term was not counted, since it could be implemented as a cascade of fast filter operations on the B-spline coefficients [87]. The fourth column ('effect $\mathcal{R}$') shows the average displacement distance between the solutions obtained with and without $\mathcal{R}$, indicating how the regularisation term affected the solution.

All methods resulted in a considerable improvement on the rigid registration. With RM, the quality of the nonrigid registration was little affected by the random subsampling strategy. Only with 256 samples the overlap and precision measures were seriously degraded. Note that the same minimum of 2048 samples was found as in the previous section, while the images considered here were almost three times larger, and the number of parameters to be optimised seven times higher. The precision of RM-2048 was somewhat better than that of GDL-4, QN-4, and NCG-4, and remained lower than the size of one voxel. The algorithms GDL-1, QN-1, QN-2, NCG-1 and NCG-2 achieved slightly better overlap and precision than RM-2048. The 'effect $\mathcal{R}$' column confirms that the solution of RM was hardly changed by adding the regularisation term.

Table 3.1: *The results for the CT chest scan application, the MR BFFE prostate scans, and the MR T1-T2 registration.*

	CT follow-up chest ($t_g \approx 220$ s.)				MR BFFE prostate ($t_g \approx 56$ s.)						MR T1-T2 prostate ($t_g \approx 9$ s.)		
	time	overlap	precision	effect $\mathcal{R}$	time	overlap	precision	effect $\mathcal{R}$	overlap*	precision*	time	precision	effect $\mathcal{R}$
	avg	avg ± sd	avg ± sd	avg ± sd	avg	avg ± sd	avg ± sd	avg ± sd	avg ± sd	avg ± sd	avg	avg ± sd	avg ± sd
	[t_g]	[%]	[mm]	[mm]	[t_g]	[%]	[mm]	[mm]	[%]	[mm]	[t_g]	[mm]	[mm]
rigid		36 ± 15	9.2 ± 7.1			37 ± 11	3.2 ± 1.0					2.9 ± 0.8	
GDL-1	700	76 ± 7	0.1 ± 0.1	1.0 ± 0.4	700	58 ± 5	0.1 ± 0.1	2.2 ± 0.9	58 ± 6	0.2 ± 0.1	100	0.6 ± 0.4	1.4 ± 0.5
GDL-2	100	75 ± 6	0.4 ± 0.1	0.9 ± 0.3	300	58 ± 5	0.2 ± 0.1	2.0 ± 0.9	58 ± 6	0.3 ± 0.2	50	0.7 ± 0.4	1.4 ± 0.5
GDL-4	10	75 ± 7	0.7 ± 0.2	0.6 ± 0.3	40	57 ± 6	0.6 ± 0.2	1.8 ± 0.7	57 ± 6	0.7 ± 0.3	10	1.1 ± 0.6	1.7 ± 0.4
GDL-8	1	71 ± 7	1.3 ± 0.3	0.6 ± 0.3	9	56 ± 6	1.6 ± 0.7	1.6 ± 0.6	55 ± 6	1.8 ± 0.7	2	1.7 ± 0.5	1.2 ± 0.3
GDL-16	0.09	60 ± 12	3.3 ± 2.6	0.9 ± 0.3	1	45 ± 6	3.3 ± 0.9	2.0 ± 0.5	43 ± 6	2.9 ± 1.0	1	3.0 ± 0.8	1.6 ± 0.7
QN-1	200	77 ± 7	0.0 ± 0.0	1.7 ± 0.8	100	58 ± 5	0.0 ± 0.0	4.0 ± 2.0	58 ± 6	0.0 ± 0.0	60	0.0 ± 0.0	4.4 ± 2.1
QN-2	40	76 ± 7	0.2 ± 0.1	1.4 ± 0.6	40	58 ± 5	0.1 ± 0.0	3.9 ± 2.0	58 ± 6	0.1 ± 0.0	20	0.4 ± 0.1	4.8 ± 2.2
QN-4	5	75 ± 7	0.7 ± 0.2	1.3 ± 0.5	8	57 ± 5	0.6 ± 0.6	3.4 ± 1.7	57 ± 6	0.5 ± 0.2	7	1.0 ± 0.6	4.9 ± 1.8
QN-8	0.5	71 ± 7	1.3 ± 0.3	1.4 ± 0.5	1	56 ± 6	1.4 ± 0.6	3.4 ± 1.2	55 ± 6	1.9 ± 0.8	2	1.9 ± 0.9	4.7 ± 1.2
QN-16	0.1	57 ± 7	2.8 ± 0.9	2.4 ± 0.8	0.2	43 ± 5	3.5 ± 0.9	4.0 ± 1.2	40 ± 8	3.2 ± 0.7	0.3	3.9 ± 1.0	4.3 ± 1.6
NCG-1	300	77 ± 7	0.1 ± 0.0	1.4 ± 0.6	200	58 ± 5	0.0 ± 0.0	3.0 ± 1.6	58 ± 6	0.0 ± 0.0	70	0.2 ± 0.2	2.8 ± 1.5
NCG-2	40	76 ± 7	0.2 ± 0.1	1.3 ± 0.6	70	58 ± 5	0.1 ± 0.1	2.8 ± 1.3	58 ± 6	0.2 ± 0.1	30	0.5 ± 0.3	2.8 ± 1.4
NCG-4	5	75 ± 7	0.7 ± 0.2	1.2 ± 0.6	10	57 ± 6	0.7 ± 0.5	2.5 ± 1.1	57 ± 6	0.6 ± 0.3	7	1.1 ± 0.6	3.4 ± 1.2
NCG-8	0.5	71 ± 8	1.4 ± 0.5	1.6 ± 0.6	2	56 ± 5	1.5 ± 0.6	2.3 ± 1.0	55 ± 6	1.7 ± 0.7	2	1.7 ± 0.7	2.7 ± 0.8
NCG-16	0.07	57 ± 9	3.4 ± 2.3	2.8 ± 2.3	0.5	46 ± 6	3.3 ± 0.9	3.0 ± 0.8	41 ± 10	3.3 ± 1.0	0.2	3.6 ± 0.9	3.2 ± 1.1
RM-all	1000	76 ± 7	0.2 ± 0.1	0.6 ± 0.3	2000	57 ± 6	0.4 ± 0.2	1.0 ± 0.4	58 ± 6	0.3 ± 0.2	3000	0.6 ± 0.4	0.9 ± 0.6
RM-10^5	30	76 ± 7	0.2 ± 0.1	0.6 ± 0.2	200	57 ± 6	0.4 ± 0.2	1.0 ± 0.4	58 ± 6	0.3 ± 0.2	700	0.7 ± 0.6	0.9 ± 0.5
RM-16384	5	76 ± 7	0.2 ± 0.1	0.6 ± 0.3	30	57 ± 6	0.4 ± 0.2	1.0 ± 0.4	58 ± 6	0.3 ± 0.2	200	0.7 ± 0.5	0.9 ± 0.5
RM-2048	0.6	75 ± 7	0.5 ± 0.1	0.8 ± 0.3	4	57 ± 6	0.4 ± 0.2	1.0 ± 0.4	58 ± 6	0.4 ± 0.2	30	0.7 ± 0.5	1.0 ± 0.5
RM-256	0.08	58 ± 5	2.6 ± 0.6	5.6 ± 1.1	0.5	57 ± 6	0.7 ± 0.4	1.1 ± 0.4	54 ± 8	1.3 ± 0.8	4	1.6 ± 0.8	2.0 ± 1.0

3.5.3 Motion between inter-fraction MR prostate scans

Prostate cancer treatment by radiation therapy requires an accurate localisation of the prostate: the tumour should receive a maximum dose, while neighbouring tissue (rectum and bladder) should be spared. The dose is delivered in several fractions. To keep track of deformations of the prostate that occur between consecutive treatment days, fast nonrigid registration is required [20, 22]. In this section we consider MR scans of the prostate, acquired with different protocols.

The images were acquired on a Philips Gyroscan NT Itera 3T MR scanner. Six volunteers were scanned on two days, 3-49 days apart. On each day, a balanced fast field echo (BFFE), a T1 and a T2 scan were taken. The BFFE scans have a dimension of $512\times512\times90$ voxels, with a voxel size of $0.49\times0.49\times1.0$ mm. The T1 and T2 have a dimension of $256\times256\times25$ voxels, with highly anisotropic voxels of $0.8\times0.8\times4.0$ mm. In the T2 the various structures within the prostate can be clearly distinguished, whereas the T1 provides a good contrast between the prostate and neighbouring tissue. The BFFE combines these characteristics and offers a good resolution, but often suffers from artefacts, caused by air in the rectum. Two types of experiments were performed: intramodality registration of BFFE scans and intermodality registration of T1 with T2 scans. In both experiments, the image acquired at the first day was selected as the moving image I_M. The image that served as a fixed image I_F was cropped to a rectangular region of interest roughly encompassing the prostate, bladder, and rectum.

All scans were first registered using an affine transform, with mutual information as the similarity measure and a four-level multiresolution strategy. After that a three-resolution nonrigid registration scheme was employed. We again limit our attention to the methods GDL, QN, NCG, and RM. For the registration of BFFE scans a B-spline control point spacing of 16 mm was used, leading to approximately 2500 parameters to be optimised. For the T1-T2 experiments a grid resolution of $30\times30\times70$ mm was used, corresponding to approximately 1000 parameters. A maximum of 2000 iterations per resolution was allowed. In all experiments we used $a = 5000$, $A = 51$, and $\alpha = 0.602$ for the gain sequence a_k. As in the previous section, a regularisation term ($\omega = 500$) appeared to be necessary, both for the BFFE-BFFE and the T1-T2 registrations. The mutual information was computed using $|L_F| = |L_M| = 32$. The BFFE experiments were also repeated with a larger number of joint histogram bins, $|L_F| = |L_M| = 64$, to investigate whether this influences the subsampling strategies. We may expect that more voxels are required to estimate the joint histogram.

For evaluation of the BFFE-BFFE registrations, manual segmentations of the prostate (including the seminal vesicles) were made by an experienced observer and approved by a radiation oncologist. We use the morphological gradient of the segmentation to compare the optimisation methods with respect to accuracy. For the T1 and T2 scans no segmentations were available. Precision is measured like in the previous section, by calculating at every voxel the distance of the deformation field to the solution obtained by QN-1.

The centre and right part of Table 3.1 present the results. The asterisk marks the results obtained with $|L_F| = |L_M| = 64$. The results of the BFFE registrations agree with those presented in the previous sections. The effect of increasing the number of bins can be observed most clearly for RM-256. Whereas with $|L_F| = |L_M| = 32$ the average overlap value equals that of RM-all, the results for $|L_F| = |L_M| = 64$ are slightly worse. With 2048 samples or more the results are comparable to those obtained with 32 joint histogram bins. For the T1-T2 experiments, the results with respect to precision followed the generally observed pattern. However, the differences in computation time were not so spectacular, since the images were rather small to begin with. For the BFFE registrations t_g was around 56 seconds, for the T1-T2 registrations t_g was around 9 seconds.

3.6 Conclusion

We have compared eight optimisation methods for nonrigid registration based on the maximisation of mutual information, in combination with a deformation field parameterised by cubic B-splines. The experiments indicate that a stochastic gradient descent technique, the Robbins-Monro process, is the preferred approach. With this method, the computation time can be extremely decreased by using a very small subset of the image to compute the derivative of the mutual information. Experiments were performed with different image modalities, image sizes, B-spline control point spacing, and number of histogram bins. In all cases the minimum number of samples required was found to be around 2000. It is very important to use a new, randomly selected subset of voxels in every iteration of the optimisation process (stochastic subsampling). If a single subset of voxels is used in all iterations (deterministic subsampling) the precision quickly deteriorates with increasing downsampling factors.

The quasi-Newton and nonlinear conjugate gradient method result in a slightly higher precision than the Robbins-Monro method, at the price of a ten to hundred times larger computation time. A point of attention when using quasi-Newton and nonlinear conjugate gradient is that a regularisation term is essential in many applications, to avoid unrealistic deformations. The gradient descent with line search improves the rate of convergence compared to the gradient descent without line search, but is slower than the quasi-Newton and conjugate gradient. The Kiefer-Wolfowitz algorithm converges reasonably fast, but suffers from a high computation time per iteration. The convergence rates of the simultaneous perturbation method and the evolution strategy are too low to make them competitive. Note that it remains to be investigated whether the conclusions can be generalised to the branch of nonparametric registration algorithms [10, 21].

A possible drawback of the Robbins-Monro method is the definition of the gain sequence a_k. The parameters involved must be tuned for each application. Some guidelines are provided in the literature on the simultaneous perturbation method [75], which work satisfactorily for the Robbins-Monro method as well, in our experience. Note that in all experiments described in this chapter the gain sequence was equal for each resolution. This indicates that the choice of the gain sequence is rather robust with respect to changes of the B-spline control point spacing and the amount of smoothing of the image. With the Robbins-Monro approach, acceleration factors of approximately 500, compared to a basic gradient descent method, can be easily achieved on many 3D nonrigid registration problems.

Chapter 4

Preconditioned Stochastic Gradient Descent

S. Klein, M. Staring, P. Andersson, and J.P.W. Pluim. Preconditioned stochastic gradient descent optimisation for monomodal image registration. *Submitted*.

This chapter presents a stochastic optimisation method for intensity-based monomodal image registration. The method is based on the Robbins-Monro (RM) stochastic gradient descent method and adds a preconditioning matrix. The derivation of the preconditioner is based on the observation that, after registration, the deformed moving image should approximately equal the fixed image. This prior knowledge allows us to approximate the Hessian at the minimum of the registration cost function, without knowing the coordinate transformation that corresponds to this minimum. The method is validated using 3D functional MRI time-series and 3D CT chest follow-up scans. The experimental results show that the preconditioned stochastic gradient descent method (PSGD) accelerates convergence and simplifies parameter selection, in comparison with RM.

4.1 Introduction

Image registration is an important technique in medical imaging applications. It refers to the process of spatially aligning images. Extensive surveys on registration methods are presented in [26, 43, 48].

We focus on intensity-based image registration with a parameterised coordinate transformation. Let $F(\boldsymbol{x}) : \Omega_F \mapsto \mathbb{R}$ and $M(\boldsymbol{x}) : \Omega_M \mapsto \mathbb{R}$ denote the *fixed* and *moving* image, respectively, with $\Omega_F, \Omega_M \subset \mathbb{R}^D$, and D the dimension of the images. Define the parameterised coordinate transformation $\boldsymbol{T}(\boldsymbol{x};\boldsymbol{\mu}) : \Omega_F \times \mathbb{R}^Q \mapsto \Omega_M$ with the parameter vector $\boldsymbol{\mu}$ of dimension Q. Examples of parametrisations include rigid, affine, and B-spline models. The aim of image registration is to find transformation parameters $\hat{\boldsymbol{\mu}}$ such that the deformed moving image $M(\boldsymbol{T}(\boldsymbol{x};\hat{\boldsymbol{\mu}}))$ resembles the fixed image $F(\boldsymbol{x})$. This can be formulated as a minimisation problem:

$$\hat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu}} \mathcal{C}(\boldsymbol{\mu}), \tag{4.1}$$

where $\mathcal{C}(\boldsymbol{\mu})$ represents the cost function. Examples of intensity-based cost functions are mean squared difference (MSD), normalised correlation (NC), and mutual information (MI).

In Chapter 3, it was demonstrated that a Robbins-Monro (RM) [62] stochastic gradient descent method efficiently solves the minimisation problem (4.1). While the presentation in Chapter 3 focussed only on nonrigid registration using MI as a cost function, in the authors' experience the RM method performs well in a much broader context, including rigid and affine transformation models, and various cost functions other than MI, such as MSD and NC. The RM method uses the following iterative scheme:

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - a_k \widetilde{\boldsymbol{g}}_k, \quad k = 0, 1, \ldots, K-1, \tag{4.2}$$

where $\widetilde{\boldsymbol{g}}_k$ denotes an approximation of the true derivative $\boldsymbol{g}_k \equiv \partial\mathcal{C}/\partial\boldsymbol{\mu}(\boldsymbol{\mu}_k)$, a_k is a scalar gain factor that determines the step size, and K is the number of iterations. The approximated derivative $\widetilde{\boldsymbol{g}}_k$ is obtained by computing $\boldsymbol{g}_k$ using only a small subset of voxels $\boldsymbol{x} \in \Omega_F$, randomly selected in every iteration k. The step size a_k is defined as a slowly decaying function of k:

$$a_k = a/(k+A)^{\alpha}, \tag{4.3}$$

with user-specified constants $a > 0$, $A \geq 1$, and $0 < \alpha \leq 1$.

The RM method has two downsides. Firstly, it is important to set proper values for the constants a, A, and α (especially the "overall gain" a). Step sizes that are too small result in slow convergence. Step sizes that are too large may lead to instabilities, possibly prohibiting convergence at all. The optimal settings are application-specific, and depend on the choice of C, the transformation model, and the image content. Secondly, the RM method is a kind of gradient descent method, which typically exposes a low rate of convergence on badly scaled cost functions, characterised by a high ($\gg 1$) condition number of the Hessian $\boldsymbol{H} \equiv \partial^2 \mathcal{C} / \partial \boldsymbol{\mu} \partial \boldsymbol{\mu}$ at $\hat{\boldsymbol{\mu}}$ [53].

To address these two issues, we propose a modification of the original RM algorithm, aimed at monomodal image registration problems.

4.2 Method

4.2.1 Preconditioned stochastic gradient descent

The use of a preconditioning matrix is a well-known technique to accelerate optimisation methods [53]. Based on the standard RM method, we define the following preconditioned stochastic gradient descent (PSGD) method:

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - a_k \boldsymbol{P} \widetilde{\boldsymbol{g}}_k, \quad k = 0, 1, \ldots, K-1, \tag{4.4}$$

where the preconditioner $\boldsymbol{P}$ is a positive definite $Q \times Q$ matrix. It serves to scale the derivative $\widetilde{\boldsymbol{g}}_k$, and should be chosen such that larger steps are taken in directions where the cost function is flat, and smaller steps in directions where the cost function has a high curvature. The theoretically optimal choice $\boldsymbol{P} = \boldsymbol{H}(\hat{\boldsymbol{\mu}})^{-1}$ makes (4.4) similar to the Newton-Raphson method. Unfortunately, since $\hat{\boldsymbol{\mu}}$ is unknown before registration, this choice of $\boldsymbol{P}$ is impossible to compute. Therefore, an approximation must be used.

4.2.2 A preconditioner for monomodal image registration

In this subsection, a preconditioning matrix for monomodal image registration problems is derived. For explanation, the MSD is used as a cost function, but the derivation is similar for any other cost function. The MSD cost function is given by:

$$\mathcal{C}(\boldsymbol{\mu}) = \frac{1}{N} \sum_{\boldsymbol{x} \in \Omega_F} \left(F(\boldsymbol{x}) - M(\boldsymbol{T}(\boldsymbol{x}; \boldsymbol{\mu})) \right)^2, \tag{4.5}$$

with N the number of $\boldsymbol{x} \in \Omega_F$. For the derivative $\boldsymbol{g}(\boldsymbol{\mu})$ and the Hessian $\boldsymbol{H}(\boldsymbol{\mu})$ we have:

$$\boldsymbol{g}(\boldsymbol{\mu}) \equiv \frac{\partial \mathcal{C}}{\partial \boldsymbol{\mu}} = -\frac{2}{N} \sum_{\boldsymbol{x} \in \Omega_F} (F - M \circ \boldsymbol{T}) \frac{\partial \boldsymbol{T}'}{\partial \boldsymbol{\mu}} \frac{\partial M}{\partial \boldsymbol{x}}, \tag{4.6}$$

$$\begin{aligned} \boldsymbol{H}(\boldsymbol{\mu}) \equiv \frac{\partial^2 \mathcal{C}}{\partial \boldsymbol{\mu} \partial \boldsymbol{\mu}} = \frac{2}{N} \sum_{\boldsymbol{x} \in \Omega_F} \Bigg[& \frac{\partial \boldsymbol{T}'}{\partial \boldsymbol{\mu}} \frac{\partial M}{\partial \boldsymbol{x}} \frac{\partial M'}{\partial \boldsymbol{x}} \frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\mu}} \\ & - (F - M \circ \boldsymbol{T}) \left(\frac{\partial^2 \boldsymbol{T}'}{\partial \boldsymbol{\mu} \partial \boldsymbol{\mu}} \frac{\partial M}{\partial \boldsymbol{x}} + \frac{\partial \boldsymbol{T}'}{\partial \boldsymbol{\mu}} \frac{\partial^2 M}{\partial \boldsymbol{x} \partial \boldsymbol{x}} \frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\mu}} \right) \Bigg], \end{aligned} \tag{4.7}$$

where the compact notation $F - M \circ \boldsymbol{T} \equiv F(\boldsymbol{x}) - M(\boldsymbol{T}(\boldsymbol{x}; \boldsymbol{\mu}))$ was introduced and the accent denotes the transpose operation. Our aim is to find an approximation $\widetilde{\boldsymbol{H}}$ to $\boldsymbol{H}(\hat{\boldsymbol{\mu}})$, whose inverse can be used as a preconditioning matrix $\boldsymbol{P}$.

When F and M are images of the same modality, we can exploit the fact that $M \circ \boldsymbol{T}$ will be approximately equal to F after successful registration: $F(\boldsymbol{x}) \approx M(\boldsymbol{T}(\boldsymbol{x};\hat{\boldsymbol{\mu}}))$, for all $\boldsymbol{x} \in \Omega_F$. Based on this approximation, the following two identities are derived:

$$F - M \circ \boldsymbol{T} = 0, \tag{4.8}$$

$$\frac{\partial M}{\partial \boldsymbol{x}} = \left[\frac{\partial \boldsymbol{T}'}{\partial \boldsymbol{x}}\right]^{-1} \frac{\partial F}{\partial \boldsymbol{x}}. \tag{4.9}$$

Substitution of (4.8) and (4.9) in (4.7) yields the following approximation $\widetilde{\boldsymbol{H}}$ of the Hessian at $\hat{\boldsymbol{\mu}}$:

$$\begin{aligned}\widetilde{\boldsymbol{H}} = \frac{2}{N} \sum_{\boldsymbol{x} \in \Omega_F} &\frac{\partial \boldsymbol{T}'}{\partial \boldsymbol{\mu}}(\boldsymbol{x};\hat{\boldsymbol{\mu}}) \left[\frac{\partial \boldsymbol{T}'}{\partial \boldsymbol{x}}(\boldsymbol{x};\hat{\boldsymbol{\mu}})\right]^{-1} \frac{\partial F}{\partial \boldsymbol{x}}(\boldsymbol{x}) \\ &\times \frac{\partial F'}{\partial \boldsymbol{x}}(\boldsymbol{x}) \left[\frac{\partial \boldsymbol{T}}{\partial \boldsymbol{x}}(\boldsymbol{x};\hat{\boldsymbol{\mu}})\right]^{-1} \frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\mu}}(\boldsymbol{x};\hat{\boldsymbol{\mu}}).\end{aligned} \tag{4.10}$$

Since $\hat{\boldsymbol{\mu}}$ is unknown, we approximate the terms $\partial \boldsymbol{T}/\partial \boldsymbol{\mu}(\boldsymbol{x};\hat{\boldsymbol{\mu}})$ and $\partial \boldsymbol{T}/\partial \boldsymbol{x}(\boldsymbol{x};\hat{\boldsymbol{\mu}})$ by $\partial \boldsymbol{T}/\partial \boldsymbol{\mu}(\boldsymbol{x};\boldsymbol{\mu}_0)$ and $\partial \boldsymbol{T}/\partial \boldsymbol{x}(\boldsymbol{x};\boldsymbol{\mu}_0)$, respectively. When the deformations are expected to be small, one may also use the approximation $\partial \boldsymbol{T}/\partial \boldsymbol{x} \approx \boldsymbol{I}$ in order to simplify the expression. In that case, we finally obtain:

$$\widetilde{\boldsymbol{H}} = \frac{2}{N} \sum_{\boldsymbol{x} \in \Omega_F} \frac{\partial \boldsymbol{T}'}{\partial \boldsymbol{\mu}}(\boldsymbol{x};\boldsymbol{\mu}_0) \frac{\partial F}{\partial \boldsymbol{x}}(\boldsymbol{x}) \frac{\partial F'}{\partial \boldsymbol{x}}(\boldsymbol{x}) \frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\mu}}(\boldsymbol{x};\boldsymbol{\mu}_0). \tag{4.11}$$

To compute the preconditioning matrix $\boldsymbol{P} = \widetilde{\boldsymbol{H}}^{-1}$ we use the eigendecomposition $\widetilde{\boldsymbol{H}} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}'$, where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_Q)$ is the diagonal eigenvalue matrix, with eigenvalues sorted such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_Q$. Define the diagonal matrix $\boldsymbol{\Lambda}^* = \mathrm{diag}(\lambda_1^*, \ldots, \lambda_Q^*)$, with elements λ_i^* defined by:

$$\lambda_i^* = \begin{cases} 1/\lambda_i & \text{if } \lambda_1/\lambda_i < \kappa_{\mathrm{MAX}} \\ \kappa_{\mathrm{MAX}}/\lambda_1 & \text{if } \lambda_1/\lambda_i \geq \kappa_{\mathrm{MAX}} \end{cases} \tag{4.12}$$

where $\kappa_{\mathrm{MAX}} \geq 1$ is a user-defined maximum condition number. The precondition matrix is then computed as $\boldsymbol{P} = \boldsymbol{V}\boldsymbol{\Lambda}^*\boldsymbol{V}'$. The maximum condition number κ_{MAX} serves as a safeguard in case of very ill-conditioned $\widetilde{\boldsymbol{H}}$, which may arise in nonrigid registration problems. In the case of rigid registration, the exact inverse can be used ($\kappa_{\mathrm{MAX}} = \infty$).

4.2.3 Step size settings

In a deterministic setting, with $\widetilde{\boldsymbol{g}}_k = \boldsymbol{g}_k$, the PSGD algorithm becomes similar to the Newton-Raphson method. The use of the inverse Hessian 'normalises' the cost function, so that a constant step size of $a_k = 1$ is the optimal choice [53]. In a stochastic setting, a constant step size is not allowed. Convergence must be enforced by letting a_k decay as a function of k [62], e.g. according to (4.3). Based on the analogy with the Newton-Raphson method in the deterministic case, we impose the starting condition $a_0 = 1$. Given user-defined A and α, it then follows that $a = A^\alpha$.

4.3 Experiments and results

The proposed PSGD method was compared to the standard RM method. Both algorithms were implemented as a part of `elastix`, an open source software package for image registration (see Chapter 2). Two applications were considered: rigid registration of 3D functional MR images (fMRI) and nonrigid registration of 3D CT chest scans.

In all registration experiments, $A = 50$, $\alpha = 0.6$, and $N = 2000$ were used (like in Chapter 3). In each iteration, the samples $x \in \Omega_F$ were selected randomly at locations off the voxel grid (not only at voxel centre coordinates). The number of iterations was set to $K = 250$.

4.3.1 Rigid registration of fMRI series

Eight fMRI time-series were available, acquired in the context of research on brain-computer interfaces (BCI) [72] in our hospital. Seven series were recorded with a 2D EPI protocol and one with a 3D PRESTO protocol. Images were obtained every 1.0-1.7 s. Each time-series consisted of $\tau \approx$ 200-400 scans. The dimensions of the images were 64×64 in-plane and 20-40 in the z-dimension. The voxel size was $4 \times 4 \times 4$ mm. To reduce noise, the images were smoothed using a Gaussian kernel with a standard deviation of half a voxel. In the BCI experiments, real-time rigid registration of each scan to the first scan is required, to compensate for motion of the head [45]. For our experiment, scans at time points $t = 0, 1, 100, 200, (300,)$ and τ were selected. All scans with $t > 0$ were registered to the scan at $t = 0$, which resulted in a total of 37 registrations. Since the head's motion was relatively small in most cases (in the order of the voxel size), the experiments were repeated with an extra initial offset (translation and rotation) to make the registration problem more challenging. The applied translations and rotations were drawn from a uniform distribution between ±8 mm and ±6°, respectively.

The parameter vector $\boldsymbol{\mu}$ was formed by $\boldsymbol{t}$ and $\boldsymbol{S\theta}$, where $\boldsymbol{t}$ is the translation vector, $\boldsymbol{\theta}$ represents the Euler angles, and $\boldsymbol{S}$ is a diagonal scaling matrix with elements:

$$s_{ii} = \left(\int_{\Omega_F} \left\| \frac{\partial \boldsymbol{T}}{\partial \theta_i} (\boldsymbol{x}; \boldsymbol{\mu}_0) \right\|^2 d\boldsymbol{x} \Big/ \int_{\Omega_F} d\boldsymbol{x} \right)^{-\frac{1}{2}} . \qquad (4.13)$$

The rotation parameters were thus scaled by the average voxel displacement caused by a small perturbation of the rotation angle. This was done to bring the values of the elements of $\boldsymbol{\mu}$ approximately in the same range, thus avoiding a very badly

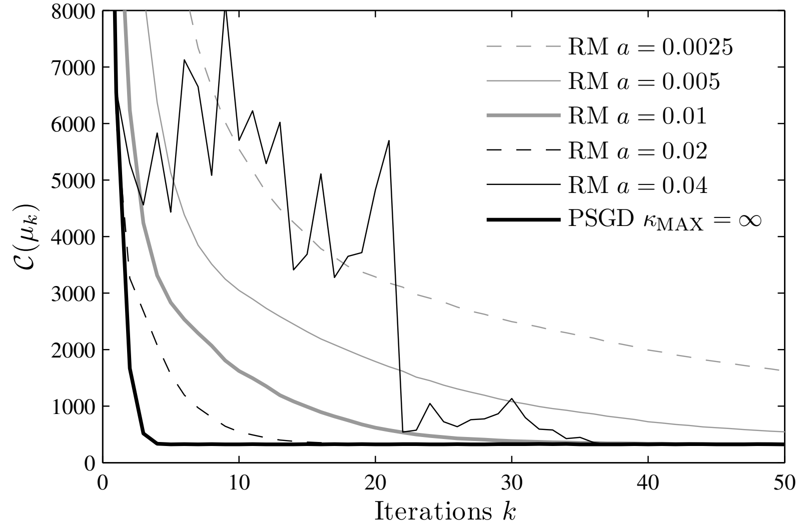


Figure 4.1: *Cost function plot for one of the fMRI experiments with additional offset.*

scaled cost function. For PSGD, the rescaling step was omitted ($\boldsymbol{S} = \boldsymbol{I}$), since it was expected that the preconditioning matrix already would take care of this. The matrix $\widetilde{\boldsymbol{H}}$ was computed according to (4.11), using $N = 50000$ randomly selected image samples. Note that $\widetilde{\boldsymbol{H}}$ needs to be computed only once, before registration. During registration $N = 2000$ samples were used, as mentioned in the introduction of this section. For computation of $\boldsymbol{P}$, $\kappa_{\mathrm{MAX}} = \infty$ was used. The RM method was tested for the range $a \in \{0.0025, 0.005, 0.01, 0.02, 0.04\}$.

For evaluation, the cost function $\mathcal{C}(\boldsymbol{\mu}_k)$ was plotted as a function of k. See Fig. 4.1 for an example. Note that $\mathcal{C}(\boldsymbol{\mu}_k)$ was calculated based on *all* voxels of the fixed image (not only the randomly sampled coordinates that were used to compute $\widetilde{\boldsymbol{g}}_k$). All graphs resulting from the experiments without the additional offset were visually inspected to determine the value of a that yielded the best convergence with RM. Based on this, RM with $a = 0.01$ was chosen as a reference standard (*ref*) for further quantitative comparison.

The differences between the coordinate transformation $\boldsymbol{T}(\boldsymbol{x}; \hat{\boldsymbol{\mu}}_{ref})$ of the reference standard and the coordinate transformations of other methods were calculated to verify whether all methods converged to the same solution. Table 4.1 presents for each method the average and standard deviation of $||\Delta \boldsymbol{T}|| \equiv ||\boldsymbol{T}(\boldsymbol{x}; \hat{\boldsymbol{\mu}}_{ref}) - \boldsymbol{T}(\boldsymbol{x}; \hat{\boldsymbol{\mu}})||$ over all $\boldsymbol{x}$ in all images. Both with and without additional offset, all differences were smaller than a voxel. The worst results were obtained using RM with $a = 0.0025$ and $a = 0.04$. To measure the rate of convergence, we counted the number of iterations before $\mathcal{C}(\boldsymbol{\mu}_k) \leq 1.01 \cdot \mathcal{C}(\hat{\boldsymbol{\mu}}_{ref})$ occurred for the first time for at least 5 consecutive iterations. The results are summarised in Table 4.1 (nr. of it.) by the average

Table 4.1: *Results of the experiments with FMRI series.*

	No additional offset		With additional offset	
	$\|\|\Delta \boldsymbol{T}\|\|$ [mm] avg ± sd	nr. of it. avg ± sd	$\|\|\Delta \boldsymbol{T}\|\|$ [mm] avg ± sd	nr. of it. avg ± sd
RM $a = 0.0025$	0.08 ± 0.15	46 ± 55	1.11 ± 1.50	209 ± 67
RM $a = 0.005$	0.03 ± 0.04	22 ± 25	0.23 ± 0.42	152 ± 86
RM $a = 0.01$ (*ref*)	0.00 ± 0.00	12 ± 11	0.01 ± 0.03	78 ± 56
RM $a = 0.02$	0.02 ± 0.02	36 ± 40	0.02 ± 0.02	58 ± 32
RM $a = 0.04$	0.37 ± 0.62	143 ± 81	0.36 ± 0.60	142 ± 82
PSGD $\kappa_{\mathrm{MAX}} = \infty$	0.07 ± 0.05	16 ± 18	0.07 ± 0.05	14 ± 12

and standard deviation over all images. The PSGD method clearly outperformed RM.

4.3.2 Nonrigid registration of CT chest scans

CT chest scans of five patients were obtained from the Department of Radiology in our institute. For each patient a baseline and a follow-up scan, taken 3-9 months later, were available. The images were downsampled by a factor of two in each direction, to reduce the burden on computer memory. The resulting voxel size was approximately $2 \times 2 \times 2$ mm. Each scan was manually cropped around the right lung, which gave images of about $120 \times 160 \times 220$ voxels. A region of interest for registration was created by dilation of a lung segmentation using a spherical kernel with radius 10.

Each follow-up scan was registered to the baseline scan using a nonrigid transformation model. A three-level multiresolution strategy was employed. The transformation was parameterised by cubic B-splines [68]. The distance between the control points was halved in each resolution level, such that at the final level the control points were spaced 20 mm in each direction. The images were smoothed using a Gaussian kernel with standard deviation of 2, 1, and 0.5 times the voxel size, for each resolution level respectively. The matrix $\widetilde{\boldsymbol{H}}$ was computed according to (4.11), using $N = 100000$ randomly selected image samples. PSGD was tested with $\kappa_{\mathrm{MAX}} \in \{1, 2, 4, 8, 16, 32\}$ and RM was tested for the range $a \in \{0.125, 0.25, 0.5, 1, 2, 4\}$.

For evaluation the same approach was followed as in the fMRI experiments. Based on visual inspection of the plotted cost function series of RM, we selected RM with $a = 0.5$ as a reference standard. Table 4.2 summarises the evaluation results. The convergence results (nr. of it.) were calculated for each resolution separately (R1-R3). The entries 250 ± 0 indicate that $\mathcal{C}(\boldsymbol{\mu}_K) > 1.01 \cdot \mathcal{C}(\hat{\boldsymbol{\mu}}_{ref})$ for all five image pairs. The numerical results in Table 4.2 indicate that PSGD achieved faster convergence than RM, although this depended on the setting of κ_{MAX}. With $\kappa_{\mathrm{MAX}} = 1$ the PSGD method boils down to an RM method with $a = A^{\alpha} / \lambda_1$. Therefore, in that

Table 4.2: *Results of the experiments with CT chest scans.*

	$\|\|\Delta\boldsymbol{T}\|\|$ [mm] avg ± sd	R1 nr. of it. avg ± sd	R2 nr. of it. avg ± sd	R3 nr. of it. avg ± sd
RM $a = 0.125$	2.36 ± 3.56	250 ± 0	250 ± 0	250 ± 0
RM $a = 0.25$	1.00 ± 1.40	250 ± 0	250 ± 0	250 ± 0
RM $a = 0.5$ (*ref*)	0.00 ± 0.00	230 ± 8	214 ± 10	211 ± 8
RM $a = 1$	1.02 ± 1.23	212 ± 52	141 ± 33	132 ± 21
RM $a = 2$	3.99 ± 7.30	250 ± 0	248 ± 5	230 ± 46
RM $a = 4$	35.04 ± 32.14	250 ± 0	250 ± 0	250 ± 0
PSGD $\kappa_{\mathrm{MAX}} = 1$	0.84 ± 1.33	250 ± 0	250 ± 0	250 ± 0
PSGD $\kappa_{\mathrm{MAX}} = 2$	0.63 ± 0.57	250 ± 0	225 ± 25	151 ± 21
PSGD $\kappa_{\mathrm{MAX}} = 4$	1.20 ± 1.27	124 ± 4	94 ± 13	102 ± 12
PSGD $\kappa_{\mathrm{MAX}} = 8$	1.84 ± 2.22	64 ± 25	52 ± 9	93 ± 16
PSGD $\kappa_{\mathrm{MAX}} = 16$	2.65 ± 3.68	45 ± 41	77 ± 97	130 ± 68
PSGD $\kappa_{\mathrm{MAX}} = 32$	3.44 ± 4.84	52 ± 72	67 ± 90	146 ± 60

case we cannot expect faster convergence than RM with optimal setting of a. Setting κ_{MAX} too high led to divergence from the reference standard's solution, as indicated by the relatively large values of $||\Delta\boldsymbol{T}||$. This can be explained by the ill-posed nature of the nonrigid registration problem. The best results were obtained with κ_{MAX} equal to 2 or 4.

4.4 Conclusion

The experiments with fMRI data show that, in the case of rigid registration, the proposed preconditioning technique has a beneficial effect on the rate of convergence. The PSGD method exhibited a rate of convergence better than RM. More importantly, the preconditioning matrix eliminated the need for manually selecting the step size parameter a and automatically compensated for the different ranges of the translation and rotation components of the transformation. In the case of nonrigid registration, the results were also promising, although the ill-posed nature of the nonrigid registration problem introduced the need for setting κ_{MAX}. The results can be interpreted in two ways. On the one hand, when setting $\kappa_{\mathrm{MAX}} = 1$, the PSGD method reduces to an RM method with automatic selection of a. On the other hand, when the user is willing to tune κ_{MAX}, faster convergence can be realised. The need for applying κ_{MAX} might be eliminated by adding a regularisation term to the registration cost function, but this would introduce another parameter: the weighting between the similarity term and the regularisation term. An interesting direction for future research might be to use the condition number of $\widetilde{\boldsymbol{H}}$ as a guidance for choosing the appropriate amount of regularisation.

The PSGD method is, just as RM, designed to work with stochastic estimates of the cost function derivatives, which leads to low computational costs per iteration (see Chapter 3). The PSGD method couples this to a good rate of convergence by using second order information of the cost function. In comparison with RM, the method is easier to use, because the selection of the step size parameter *a* has been automated.

Chapter 5

Adaptive Stochastic Gradient Descent

S. Klein, J.P.W. Pluim, M. Staring, and M.A. Viergever. Adaptive stochastic gradient descent optimisation for image registration. *International Journal of Computer Vision*, in press.

This chapter presents a stochastic gradient descent optimisation method for image registration with adaptive step size prediction. The method is based on the theoretical work by Plakhov and Cruz [58]. Our main methodological contribution is the derivation of an image-driven mechanism to select proper values for the most important free parameters of the method. The selection mechanism employs general characteristics of the cost functions that commonly occur in intensity-based image registration. Also, the theoretical convergence conditions of the optimisation method are taken into account. The proposed adaptive stochastic gradient descent (ASGD) method is compared to a standard, non-adaptive Robbins-Monro (RM) algorithm. Both ASGD and RM employ a stochastic subsampling technique to accelerate the optimisation process. Registration experiments were performed on 3D CT and MR data of the head, lungs, and prostate, using various similarity measures and transformation models. The results indicate that ASGD is robust to these variations in the registration framework and is less sensitive to the settings of the user-defined parameters than RM. The main disadvantage of RM is the need for a predetermined step size function. The ASGD method provides a solution for that issue.

5.1 Introduction

Image registration is a frequently used technique in the fields of remote sensing and medical imaging. Given a pair of images, image registration is the task of finding a coordinate transformation that spatially aligns the two images. Extensive surveys of registration methods can be found in the literature [43, 47, 99]. We focus on intensity-based registration methods, using a parameterised coordinate transformation.

Intensity-based image registration is usually treated as a nonlinear optimisation problem. Define the fixed image $F(\boldsymbol{x}) : \Omega_F \subset \mathbb{R}^D \mapsto \mathbb{R}$, the moving image $M(\boldsymbol{x}) : \Omega_M \subset \mathbb{R}^D \mapsto \mathbb{R}$, and a parameterised coordinate transformation $\boldsymbol{T}(\boldsymbol{x}, \boldsymbol{\mu}) : \Omega_F \times \mathbb{R}^P \mapsto \Omega_M$, where $\boldsymbol{\mu} \in \mathbb{R}^P$ represents the vector of transformation parameters. The following minimisation problem is considered:

$$\hat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu}} \mathcal{C}\left(F, M \circ \boldsymbol{T}\right), \tag{5.1}$$

where $\mathcal{C}$ is the cost function (or "similarity measure") that measures the similarity of the fixed image and the deformed moving image. The solution $\hat{\boldsymbol{\mu}}$ is the parameter vector that minimises that cost function. Henceforth, we use the short notation $\mathcal{C}(\boldsymbol{\mu}) \equiv \mathcal{C}\left(F, M \circ \boldsymbol{T}\right)$.

In Chapter 3 it has been shown that a Robbins-Monro (RM) stochastic gradient descent method [36, 62] is in many applications the best choice for solving the minimisation problem (5.1). The method uses the following iterative scheme:

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - \gamma_k \widetilde{\boldsymbol{g}}_k, \quad k = 0, 1, \ldots, K, \tag{5.2}$$

$$\widetilde{\boldsymbol{g}}_k = \boldsymbol{g}(\boldsymbol{\mu}_k) + \boldsymbol{\varepsilon}_k, \tag{5.3}$$

where $\widetilde{\boldsymbol{g}}_k$ denotes an approximation of the true derivative $\boldsymbol{g} \equiv \partial\mathcal{C}/\partial\boldsymbol{\mu}$ at $\boldsymbol{\mu}_k$, and $\boldsymbol{\varepsilon}_k$ is the approximation error. If $\boldsymbol{\varepsilon}_k = \boldsymbol{0}$, Eq. (5.2) boils down to a common, deterministic gradient descent method. The approximation of $\boldsymbol{g}$ is realised by computing $\boldsymbol{g}$ using not all voxels, but only a small subset of voxels, randomly selected in every iteration. In this way, the computational costs per iteration are greatly reduced, while convergence properties are still similar to those obtained by deterministic gradient descent. The scalar gain factor γ_k, the "step size", is determined by a predefined decaying function of the iteration number k. An often used choice is:

$$\gamma_k \equiv \gamma(k) = a/(k + A)^{\alpha}, \tag{5.4}$$

with user-specified constants $a > 0$, $A \geq 1$, and $0 < \alpha \leq 1$. A choice of $\alpha = 1$ gives a theoretically optimum rate of convergence when $k \to \infty$ [36]. In practice, the algorithm is stopped after a specified maximum number of iterations, and, therefore, it sometimes makes sense to choose $\alpha < 1$, which causes the step size to decay less fast. The need for setting a, A, and α complicates the usage of RM for image registration. The factor a is especially difficult, since it has no unit, and heavily depends on the choice of the cost function. For example, when we multiply $\mathcal{C}$ by an arbitrary constant c, the value of a would need to be divided by c in order to get the same sequence $\{\boldsymbol{\mu}_k\}$. When a is set too small, the RM method suffers from slow convergence. When a is set too large, the process may become unstable.

The present study concerns a stochastic optimisation method with adaptive step size prediction: adaptive stochastic gradient descent (ASGD). The mechanism to adapt the step size γ_k is based on the inner product of the gradient $\widetilde{\boldsymbol{g}}_k$ and the previous gradient $\widetilde{\boldsymbol{g}}_{k-1}$. Intuitively, if the gradients in two consecutive iterations point in (almost) the same direction, it is expected that larger steps can be taken. If the gradients point in opposite directions, the step size is reduced. The theoretical convergence properties of the method in one-dimensional ($P = 1$) optimisation problems were studied by Plakhov and Cruz [58]. Cruz [13] extended the analysis to multidimensional ($P > 1$) problems. Some numerical experiments are described in [14], using artificial test functions with $\boldsymbol{\varepsilon}_k$ generated according to a normal distribution. Only two cases ($P = 1$ and $P = 2$) were investigated. No other applications of the method were found in the literature.

Using the theoretical convergence conditions given in [13], we derive an image-driven selection mechanism for the free parameters of the method. The derivation is based on general characteristics of the cost functions that commonly occur in intensity-based image registration problems. A key result is the replacement of a by a new user-defined parameter, δ, which has a more intuitive meaning and is constructed to be independent of the choice of $\mathcal{C}$. The method is validated on several registration problems, with different image modalities, similarity measures, and transformation models, with P ranging from 6 to 4000.

5.2 Method

First, in Sec. 5.2.1, the basic ASGD method is explained and a summary is given of the theoretical convergence results. After that, in Sec. 5.2.2, we describe the first steps towards application of ASGD in image registration. A procedure to set the free parameters of ASGD is derived in Secs. 5.2.3-5.2.5. Section 5.2.6 gives an overview of the entire algorithm.

5.2.1 Summary of ASGD

In [13] the ASGD method is presented in the context of a general multidimensional root-finding problem[†]: find $\hat{\boldsymbol{\mu}}$ such that $\boldsymbol{\varphi}(\hat{\boldsymbol{\mu}}) = \mathbf{0}$, for some function $\boldsymbol{\varphi}(\boldsymbol{\mu}) : \mathbb{R}^P \mapsto \mathbb{R}^P$. Our minimisation problem is a specific case of this, where $\boldsymbol{\varphi}$ equals $\boldsymbol{g} \equiv \partial\mathcal{C}/\partial\boldsymbol{\mu}$. The ASGD algorithm is then defined as:

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - \gamma(t_k)\widetilde{\boldsymbol{g}}_k, \quad k = 0, 1, \dots, K, \tag{5.5}$$

$$t_{k+1} = \left[t_k + f(-\widetilde{\boldsymbol{g}}_k'\widetilde{\boldsymbol{g}}_{k-1})\right]^+, \tag{5.6}$$

[†] Note that our notation is somewhat different from [13].

where the accent denotes the transpose operation, $[x]^+$ means $\max(x,0)$, f denotes a sigmoid function, and $\boldsymbol{\mu}_0$, t_0 and t_1 are user-defined initial conditions. For the γ function, the same definition as in (5.4) can be used. However, in ASGD, the γ function is not evaluated at the iteration number k, as in (5.2), but at the "time" t_k. The time is adapted depending on the inner product of the gradient $\widetilde{\boldsymbol{g}}_k$ and the previous gradient $\widetilde{\boldsymbol{g}}_{k-1}$. If the gradients in two consecutive steps point in the same direction, the inner product is positive, and therefore the time is reduced, which leads to a larger step size $\gamma(t_{k+1})$, since γ is a monotone decreasing function. In this way, the ASGD method implements an adaptive step size mechanism. Note that if we would use $f(x) = 1$, the original RM method is obtained.

The article by Cruz [13] provides a proof of "almost-sure" convergence and a proof of asymptotical normality. The proof of almost-sure convergence implies that

$$\lim_{k\to\infty} \boldsymbol{\mu}_k = \hat{\boldsymbol{\mu}}, \tag{5.7}$$

"with probability 1". The proof of asymptotical normality tells us something about the rate of convergence:

$$\sqrt{k}(\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}}) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{V}), \tag{5.8}$$

where $\xrightarrow{d}$ indicates convergence in distribution and $\mathcal{N}(\boldsymbol{0}, \boldsymbol{V})$ denotes a multivariate normal distribution with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{V}$. To prove the convergence and asymptotical normality, five sets of assumptions are required. The assumptions impose conditions on γ and f, depending on characteristics of the cost function $\mathcal{C}$ and the distribution of gradient approximation errors $\boldsymbol{\varepsilon}_k$.

5.2.2 Application of ASGD

To apply the ASGD method in practice, we have to specify the γ and f functions. They should be chosen such that the theoretical convergence conditions given in [13] are satisfied.

For the step size function γ we choose the following expression:

$$\gamma(t) = a/(t + A), \tag{5.9}$$

with $a > 0$ and $A \geq 1$. Compared to (5.4) the α term is omitted, i.e. $\alpha = 1$, which is the theoretically optimal setting [36]. For f we define a general sigmoid shape with $f(0) = 0$:

$$f(x) = f_{\text{MIN}} + \frac{f_{\text{MAX}} - f_{\text{MIN}}}{1 - (f_{\text{MAX}}/f_{\text{MIN}})e^{-x/\omega}}, \tag{5.10}$$

with $f_{\text{MAX}} > 0$, $f_{\text{MIN}} < 0$, and $\omega > 0$. Examples of f are shown in Fig. 5.1. If $\omega \downarrow 0$, the sigmoid approaches a step function.

The ASGD algorithm still requires setting a and A. Moreover, the expression for the sigmoid function f introduces three new parameters: f_{MAX}, f_{MIN}, and ω. Yet, we

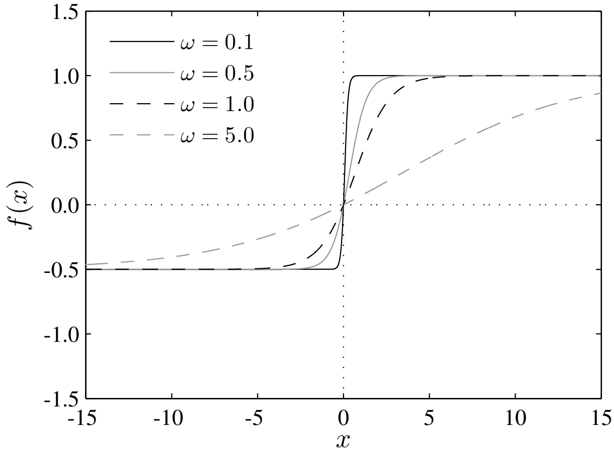


Figure 5.1: *Examples of the sigmoid function f, with $f_{\text{MAX}} = 1$ and $f_{\text{MIN}} = -0.5$.*

expect that the adaptive step size mechanism makes the algorithm robust for wider ranges of a and A, compared with RM.

As mentioned in Sec. 5.2.1, five sets of assumptions are used in [13] to prove convergence and asymptotical normality of the ASGD algorithm. The assumptions impose conditions on γ and f, and are thus important for determining proper values for a, A, f_{MIN}, f_{MAX}, and ω. We now study the assumptions after substitution of the above choices for γ and f. Like in [13] the sets of assumptions needed to prove convergence are numbered B1-B4. The set of assumptions used to prove asymptotical normality is called B5. In comparison with [13], some conditions have been simplified using $\boldsymbol{\varphi} = \partial \mathcal{C} / \partial \boldsymbol{\mu} \equiv \boldsymbol{g}$ (see Sec. 5.2.1). Also, technical details that are not relevant for this work are omitted. Our comments on the assumptions are written in *italic*.

Assumption B1 - properties of ε_k

The approximation errors $\boldsymbol{\varepsilon}_k$ are independent identically distributed random vectors with zero mean $\mathrm{E}\boldsymbol{\varepsilon}_k = 0$ and finite covariance matrix $\boldsymbol{\Sigma} \equiv \mathrm{Var}\boldsymbol{\varepsilon}_k$.

Based on characteristics of the cost function $\mathcal{C}$, we postulate in Sec. 5.2.3 that $\boldsymbol{\varepsilon}_k$ has a normal distribution: $\boldsymbol{\varepsilon}_k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$.

Assumptions B2 - properties of $\gamma(t)$

1. The gain function $\gamma(t)$ is a positive monotone decreasing function defined on $[0, \infty)$. Consequently, $\gamma(0)$ is the maximum gain factor.

2. $\int_0^\infty \gamma(t) dt = \infty$.

3. $\int_0^\infty [\gamma(t)]^2 dt < \infty$.

With $\gamma(t)$ defined by (5.9) it is easily verified that these assumptions are satisfied and that $\gamma(0) = a/A$.

Assumptions B3 - conditions depending on $\mathcal{C}$

1. Provided that

   a) the function $\mathcal{C}(\boldsymbol{\mu})$ has no other extrema than $\hat{\boldsymbol{\mu}}$,

   b) $\mathcal{C}(\boldsymbol{\mu})$ is continuous and twice differentiable everywhere,

   c) there exists a constant $\lambda > 0$ such that the maximum eigenvalue of the Hessian $\boldsymbol{H} \equiv \partial^2 \mathcal{C} / \partial \boldsymbol{\mu} \partial \boldsymbol{\mu}$ is smaller than or equal to λ for all $\boldsymbol{\mu}$,

   the minimisation problem (5.1) can be solved with the following deterministic gradient descent method:

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - \hat{\gamma} \boldsymbol{g}(\boldsymbol{\mu}_k), \tag{5.11}$$

   for each $\hat{\gamma} < \gamma(0)$, and for each $\boldsymbol{\mu}_0$.

   Provided that B3.1(a)-(c) indeed hold, the choice $\gamma(0) = 2/\lambda$ satisfies the last condition [71]. This assumption thus relates the maximum step size $\gamma(0)$ to the Hessian of the cost function.

2. There exist $R > 0$ and $\beta_0 > 0$ such that[†]

$$||\boldsymbol{g}(\boldsymbol{\mu})||^2 \geq \tfrac{1}{2}\gamma(0)\lambda \left(||\boldsymbol{g}(\boldsymbol{\mu})||^2 + \operatorname{tr}(\boldsymbol{\Sigma}) \right) + \beta_0, \tag{5.12}$$

   for all $\boldsymbol{\mu}$ that satisfy $||\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}|| \geq R$.

   This condition relates the maximum step size $\gamma(0)$ to the covariance matrix $\boldsymbol{\Sigma}$ of the approximation errors. In Sec. 5.2.4, we use assumptions B3.1 and B3.2 to choose a value for a.

Assumptions B4 - properties of $f(x)$

1. $f(x) : \mathbb{R} \mapsto \mathbb{R}$ is a monotone increasing, continuous and bounded function, for which:

$$f_{\text{MAX}} = \lim_{x \to +\infty} f(x) > 0 \quad \text{and} \quad f_{\text{MIN}} = \lim_{x \to -\infty} f(x) \tag{5.13}$$

   The expression for $f(x)$ defined in (5.10) has been constructed such that B4.1 is satisfied.

2. Define $E_0 \equiv \mathrm{E} f(\boldsymbol{\varepsilon}_k' \boldsymbol{\varepsilon}_{k-1})$. The constant E_0 must be positive.

   The condition $E_0 > 0$ is satisfied when $f(x) > -f(-x)$ for all $x \neq 0$, provided that $\boldsymbol{\Sigma} \neq \boldsymbol{0}$. Combined with (5.10), this imposes that $f_{\text{MAX}} > -f_{\text{MIN}}$. Furthermore, if $\boldsymbol{\varepsilon}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ and $\omega \downarrow 0$, then $E_0 \uparrow \frac{1}{2}(f_{\text{MAX}} + f_{\text{MIN}})$. In Sec. 5.2.5 this is used to choose values for f_{MAX} and f_{MIN}. Also, a value for ω is determined, such that indeed $E_0 \approx \frac{1}{2}(f_{\text{MAX}} + f_{\text{MIN}})$.

[†] $\operatorname{tr}(\cdot)$ stands for the matrix trace.

Assumptions B5 - asymptotic normality

The following conditions are used to prove asymptotic normality:

1. $\gamma(t) = 1/t$.
2. Define the matrix $\boldsymbol{W}$:

$$\boldsymbol{W} = \frac{1}{2}\boldsymbol{I} - \frac{1}{E_0}\boldsymbol{H}(\hat{\boldsymbol{\mu}}), \tag{5.14}$$

with $\boldsymbol{I}$ the identity matrix. All eigenvalues of $\boldsymbol{W}$ must be negative.

3. $f(x)$ is a continuous and differentiable function.

Assumption B5.3 is obviously satisfied when $\omega \neq 0$. Our choice for $\gamma(t)$ breaks with Assumption B5.1. However, the proof of asymptotic normality can be easily extended to take our choice of $\gamma(t)$ into account. The first two assumptions are then modified to:

1. $\gamma(t) = a/(t + A)$.
2. Define the matrix $\boldsymbol{W}$:

$$\boldsymbol{W} = \frac{1}{2}\boldsymbol{I} - \frac{a}{E_0}\boldsymbol{H}(\hat{\boldsymbol{\mu}}), \tag{5.15}$$

with $\boldsymbol{I}$ the identity matrix. All eigenvalues of $\boldsymbol{W}$ must be negative.

Assumptions B4.2 and B5.2 are used in Sec. 5.2.5 to choose a value for f_{MAX} and f_{MIN}.

In the following subsections, estimates are derived for the distributions of $\boldsymbol{g}$, $\widetilde{\boldsymbol{g}}_k$, $\boldsymbol{\varepsilon}_k$, and the voxel displacements between two iterations. Based on these results and some of the assumptions B1-B5 mentioned above, settings for a, f_{MIN}, f_{MAX}, and ω are proposed. The value of A is left unspecified. The parameter a is replaced by a new user-defined parameter δ, which, unlike a, has a unit (mm), and an intuitive meaning. Also, it is constructed to be independent of the choice of $\mathcal{C}$.

5.2.3 Distribution estimates

In this subsection we devise expressions for the distributions of $\boldsymbol{g}$, $\widetilde{\boldsymbol{g}}_k$ and $\boldsymbol{\varepsilon}_k$, based on the characteristics of the cost function in image registration problems. Using the distribution of $\boldsymbol{g}$, the distribution of voxel displacements per iteration of a deterministic gradient descent process is calculated. The results of this subsection are needed in Secs. 5.2.4 and 5.2.5.

In image registration, the cost function usually takes the following form:

$$\mathcal{C}(\boldsymbol{\mu}) = \Psi\left(\frac{1}{|\Omega_F^V|}\sum_{\boldsymbol{x}_i \in \Omega_F^V} \xi\Big(F(\boldsymbol{x}_i), M(\boldsymbol{T}(\boldsymbol{x}_i, \boldsymbol{\mu}))\Big)\right), \tag{5.16}$$

with $\Psi(u) : \Xi \mapsto \mathbb{R}$ and $\xi(u, v) : \mathbb{R} \times \mathbb{R} \mapsto \Xi$ continuous, differentiable functions, $\Omega_F^V \subset \Omega_F$ the discrete set of voxel coordinates $\boldsymbol{x}_i$ of the fixed image, and $|\Omega_F^V|$ the cardinality of this set. The domain Ξ may be simply equal to $\mathbb{R}$, but may also be of a multidimensional nature: $\mathbb{R}^P$ or $\mathbb{R}^{P \times Q}$, for example. An example that is covered by (5.16) is the sum of squared differences: $\Xi = \mathbb{R}$, $\Psi(u) = u$, $\xi(u,v) = (u-v)^2$. Another example is mutual information [85], for which:

$$\Xi = \mathbb{R}^{P \times Q}, \tag{5.17}$$

$$\Psi(u) = \sum_{p=1}^{P} \sum_{q=1}^{Q} u_{pq} \log \frac{u_{pq}}{\left(\sum_p u_{pq}\right)\left(\sum_q u_{pq}\right)}, \tag{5.18}$$

$$\xi_{pq}(u,v) = \beta(p-u)\beta(q-v), \tag{5.19}$$

with $P \times Q$ the joint histogram size, and $\beta(u) : \mathbb{R} \mapsto \mathbb{R}$ a Parzen window function.

We now take the derivative of (5.16). For clarity of notation we consider the case $\Xi = \mathbb{R}$:

$$\boldsymbol{g} \equiv \frac{\partial \mathcal{C}}{\partial \boldsymbol{\mu}} = \frac{1}{|\Omega_F^V|} \sum_{\boldsymbol{x}_i \in \Omega_F^V} \frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\mu}}' \frac{\partial M}{\partial \boldsymbol{x}} \frac{\partial \xi}{\partial v} \frac{\partial \Psi}{\partial u}. \tag{5.20}$$

We would like to estimate the distribution of $\boldsymbol{g}$ in a neighbourhood $\Upsilon \subset \mathbb{R}^P$ around $\hat{\boldsymbol{\mu}}$, containing $\boldsymbol{\mu}_0$. The idea is that this distribution predicts the gradients that will be measured during optimisation. The following two assumptions are needed:

Assumption A1 - $\partial T/\partial \mu$ is independent of μ
For each $\boldsymbol{x}_i \in \Omega_F^V$ the following holds:

$$\boldsymbol{J}_i \equiv \frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\mu}}(\boldsymbol{x}_i, \boldsymbol{\mu}_0) = \frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\mu}}(\boldsymbol{x}_i, \boldsymbol{\mu}), \quad \forall \boldsymbol{\mu} \in \Upsilon. \tag{5.21}$$

This assumption holds when the transformation model is parameterised such that $\partial^2 \boldsymbol{T}/\partial\boldsymbol{\mu}\partial\boldsymbol{\mu} = \boldsymbol{0}$. The B-spline transformation [68] is an example of such a parametrisation. Also an affine transformation, parameterised by the affine matrix elements, satisfies the assumption. For a rigid transformation parameterised by Euler angles the assumption is violated, since $\boldsymbol{T}$ then becomes a nonlinear function of $\boldsymbol{\mu}$, but it holds approximately if Υ *is not too large.*

Assumption A2 - distribution of z_i
Based on (5.20), define:

$$\boldsymbol{z}_i \equiv \frac{\partial M}{\partial \boldsymbol{x}} \frac{\partial \xi}{\partial v} \frac{\partial \Psi}{\partial u}. \tag{5.22}$$

Then, $\{\boldsymbol{z}_i\}$ are mutually independent random vectors, identically distributed according to:

$$\boldsymbol{z}_i \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}\right), \tag{5.23}$$

with σ some constant.
This assumption is a simplification of reality. Any results based on this assumption must be validated.

Combining (5.20)-(5.23) gives us an estimate of the distribution of $\boldsymbol{g}$:

$$\boldsymbol{g} = \frac{1}{|\Omega_F^V|} \sum_{\boldsymbol{x}_i \in \Omega_F^V} \boldsymbol{J}_i' z_i \sim \mathcal{N}\left(\boldsymbol{0}, \frac{\sigma^2}{|\Omega_F^V|^2} \sum_{\boldsymbol{x}_i \in \Omega_F^V} \boldsymbol{J}_i' \boldsymbol{J}_i\right) = \mathcal{N}\left(\boldsymbol{0}, \frac{\sigma^2}{|\Omega_F^V|} \boldsymbol{C}\right), \tag{5.24}$$

where we introduced:

$$\boldsymbol{C} \equiv \frac{1}{|\Omega_F^V|} \sum_{\boldsymbol{x}_i \in \Omega_F^V} \boldsymbol{J}_i' \boldsymbol{J}_i. \tag{5.25}$$

The same approach can be followed for the approximated derivative $\widetilde{\boldsymbol{g}}_k$. Approximation is realised by stochastic subsampling:

$$\widetilde{\boldsymbol{g}}_k = \frac{1}{|S_k|} \sum_{\boldsymbol{x}_i \in S_k} \frac{\partial \boldsymbol{T}'}{\partial \boldsymbol{\mu}} \frac{\partial M}{\partial \boldsymbol{x}} \frac{\partial \xi}{\partial v} \frac{\partial \Psi}{\partial u}, \tag{5.26}$$

with $S_k \subset \Omega_F^V$ a set of samples, randomly selected in every iteration k. The distribution for $\widetilde{\boldsymbol{g}}_k$ is estimated in the same way as above:

$$\widetilde{\boldsymbol{g}}_k \sim \mathcal{N}\left(\boldsymbol{0}, \frac{\sigma^2}{|S_k|^2} \sum_{\boldsymbol{x}_i \in S_k} \boldsymbol{J}_i' \boldsymbol{J}_i\right). \tag{5.27}$$

The following approximation is proposed:

$$\frac{1}{|S_k|} \sum_{\boldsymbol{x}_i \in S_k} \boldsymbol{J}_i' \boldsymbol{J}_i \approx \frac{1}{|\Omega_F^V|} \sum_{\boldsymbol{x}_i \in \Omega_F^V} \boldsymbol{J}_i' \boldsymbol{J}_i. \tag{5.28}$$

The approximation becomes more accurate for increasing $|S_k|$, and when $\boldsymbol{J}_i$ varies more gradually over the image domain Ω_F. Using this approximation, the expression for the distribution of $\widetilde{\boldsymbol{g}}_k$ becomes:

$$\widetilde{\boldsymbol{g}}_k \sim \mathcal{N}\left(\boldsymbol{0}, \frac{\sigma^2}{|S_k|} \boldsymbol{C}\right). \tag{5.29}$$

The distribution of the approximation errors $\boldsymbol{\varepsilon}_k = \boldsymbol{g} - \widetilde{\boldsymbol{g}}_k$ is computed in a similar way, by subtracting (5.26) from (5.20), and using the approximation (5.28):

$$\boldsymbol{\varepsilon}_k \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2 \left(\frac{1}{|\Omega_F^V|} - \frac{1}{|S_k|}\right) \boldsymbol{C}\right). \tag{5.30}$$

Note that when the number of samples $|S_k|$ is independent of k, the distributions of $\widetilde{\boldsymbol{g}}_k$ and $\boldsymbol{\varepsilon}_k$ are also independent of k.

We turn our attention to Assumption A2. The assumptions states that $\boldsymbol{z}_i$ are independent random variables. In images, the assumption of independency between neighbouring voxels $\boldsymbol{x}_i$ is usually not satisfied. Consequently, the corresponding $\boldsymbol{z}_i$ may also be related. The impact of this on the distribution of $\boldsymbol{g}$, see (5.24), can be demonstrated by an imaginary experiment. Suppose we have experimentally estimated the distribution of $\boldsymbol{g}$ in some region Y. Then, we resample the fixed image $F(\boldsymbol{x})$ on a twice as dense grid, using for example linear interpolation to interpolate between voxels, and repeat the experiment. Intuitively, we would not expect a different distribution of $\boldsymbol{g}$. However, the number of voxels in the fixed image, $|\Omega_F^V|$, has increased with a factor 2^D, with D the dimension of the fixed image. According to (5.24), the variance of the distribution should therefore be divided by a factor 2^D, which is clearly wrong. We must conclude that the dependency of the variance on the number of voxels only holds when the z_i are truly independent. Since this is hard to verify, we propose to use the following distribution estimates, instead of (5.24), (5.29), and (5.30):

$$\boldsymbol{g} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma_1^2 \boldsymbol{C}\right), \tag{5.31}$$

$$\widetilde{\boldsymbol{g}}_k \sim \mathcal{N}\left(\boldsymbol{0}, \sigma_2^2 \boldsymbol{C}\right), \tag{5.32}$$

$$\boldsymbol{\varepsilon}_k \sim \mathcal{N}\left(\boldsymbol{0}, \sigma_3^2 \boldsymbol{C}\right) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \tag{5.33}$$

with σ_1, σ_2, and σ_3 unknown scalar constants, unrelated to each other, which will be experimentally determined. The last equality refers to the comment on Assumption B1. To estimate the constants σ_i, we perform N evaluations of $\boldsymbol{g}$, $\widetilde{\boldsymbol{g}}_k$, and $\boldsymbol{\varepsilon}_k = \boldsymbol{g} - \widetilde{\boldsymbol{g}}_k$, and fit σ_i such that the empirical average vector magnitudes equal the theoretical expectations of the vector magnitudes. For example, σ_1 is determined such that:

$$\frac{1}{N}\sum_{n=1}^{N} ||\boldsymbol{g}(\boldsymbol{\mu}_n)||^2 = \sigma_1^2 \mathrm{tr}(\boldsymbol{C}), \tag{5.34}$$

where the right-hand side equals $\mathrm{E}||\boldsymbol{g}||^2$, in accordance with (5.31). The $\boldsymbol{\mu}_n$ vectors are randomly sampled around $\boldsymbol{\mu}_0$, using a normal distribution with diagonal covariance matrix:

$$\boldsymbol{\mu}_n \sim \mathcal{N}(\boldsymbol{\mu}_0, \sigma_4^2 \boldsymbol{I}), \tag{5.35}$$

where σ_4 is a scalar constant, chosen such that the voxel displacements $||\boldsymbol{T}(\boldsymbol{x}_j, \boldsymbol{\mu}_n) - \boldsymbol{T}(\boldsymbol{x}_j, \boldsymbol{\mu}_0)||$ caused by the parameter change from $\boldsymbol{\mu}_0$ to $\boldsymbol{\mu}_n$ remain with high probability (≈ 0.95) below a user-defined value δ. The exact procedure is explained at the end of Sec. 5.2.4. The user-defined constant N, introduced in (5.34), should be chosen high enough such that $\frac{1}{N}\sum_n ||\boldsymbol{g}(\boldsymbol{\mu}_n)||^2$ is a good estimate of the true expectation

$\mathrm{E}||\boldsymbol{g}||^2$. When $\boldsymbol{C}$ equals the identity matrix, the average squared gradient magnitude $\frac{1}{N}\sum_n ||\boldsymbol{g}(\boldsymbol{\mu}_n)||^2$ has a χ^2_{NP} distribution. The ratio between the standard deviation and the expectation of a χ^2_{NP} distribution equals $\sqrt{2}/\sqrt{NP}$. We can thus expect that, with increasing P, N can be lowered. For arbitrary $\boldsymbol{C}$, the ratio between standard deviation and expectation can be shown to have an upper bound of $\sqrt{2/N}$. From this, it is clear that a value $N \approx 10$ is a reasonable choice.

Having estimated the distribution of $\boldsymbol{g}$, we can calculate the distribution of voxel displacements per iteration of a deterministic gradient descent process. The deterministic gradient descent procedure mentioned in Assumption B3.1, Eq. (5.11), is considered: $\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - \hat{\gamma}\boldsymbol{g}(\boldsymbol{\mu}_k)$. The displacement $\boldsymbol{d}_k$ of voxel $\boldsymbol{x}_j$ between iteration k and $k+1$ is defined by:

$$\boldsymbol{d}_k(\boldsymbol{x}_j) \equiv \boldsymbol{T}(\boldsymbol{x}_j, \boldsymbol{\mu}_{k+1}) - \boldsymbol{T}(\boldsymbol{x}_j, \boldsymbol{\mu}_k). \tag{5.36}$$

Our goal is to estimate the distribution of $\boldsymbol{d}_k(\boldsymbol{x}_j)$ for some $\boldsymbol{\mu}_k \in \mathrm{Y}$. This result is used in Sec. 5.2.4 to estimate λ (see Assumption B3.1), which is used to select a such that Assumptions B3 are satisfied. According to the Taylor expansion of $\boldsymbol{T}$ around $\boldsymbol{\mu}_k$:

$$\boldsymbol{d}_k(\boldsymbol{x}_j) \approx \frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\mu}}(\boldsymbol{x}_j, \boldsymbol{\mu}_k) \cdot (\boldsymbol{\mu}_{k+1} - \boldsymbol{\mu}_k) = \boldsymbol{J}_j\left(\boldsymbol{\mu}_{k+1} - \boldsymbol{\mu}_k\right), \tag{5.37}$$

where the last equality follows from Assumption A1. Substitution of (5.11) gives:

$$\boldsymbol{d}_k(\boldsymbol{x}_j) \approx -\hat{\gamma}\boldsymbol{J}_j\boldsymbol{g}(\boldsymbol{\mu}_k). \tag{5.38}$$

Using (5.31) we obtain:

$$\boldsymbol{d}_k(\boldsymbol{x}_j) \sim \mathcal{N}\left(\boldsymbol{0}, \hat{\gamma}^2\sigma_1^2\boldsymbol{J}_j\boldsymbol{C}\boldsymbol{J}_j'\right). \tag{5.39}$$

Note that the estimated distribution of $\boldsymbol{d}_k(\boldsymbol{x}_j)$ is independent of k.

The distribution estimates that have been derived in this subsection are used in the following subsections. In Sec. 5.2.4, Eqs. (5.31), (5.33), and (5.39) are used to select a. Equation (5.33) is used also in Sec. 5.2.5 to select ω.

5.2.4 Selection of a

In this subsection an appropriate value of a is estimated, using Assumptions B3 and Eqs. (5.31), (5.33), and (5.39). The value of A is considered a user-defined constant. The method consists of two steps. First, a deterministic gradient descent method is considered. The maximum value of a that still ensures convergence is estimated, based on Assumption B3.1, Eq. (5.39), and an additional user input: the maximum allowed voxel displacement δ. After that, Assumption B3.2 is combined with Eqs. (5.31) and (5.33), to derive an expression for a that takes the stochastic approximation errors into account.

As mentioned in Sec. 5.2.2, Assumption B3.1, the maximum value for $\gamma(0) = a/A$ that ensures convergence of the deterministic gradient process (5.11) equals

$2/\lambda$, provided that conditions B3.1(a)-(c) hold. Condition B3.1(a) is often not satisfied in image registration problems. This is a general problem of image registration, which will not be further addressed in this chapter. Henceforth, we simply assume that $\boldsymbol{\mu}_0$ is chosen within the capture range of the desired local minimum $\hat{\boldsymbol{\mu}}$. The value of λ, which is defined by condition B3.1(c), is generally unknown. We propose to estimate λ based on an additional user input parameter δ: the maximum allowed magnitude of the voxel displacements $\boldsymbol{d}_k(\boldsymbol{x}_j)$. The problem becomes thus to compute a λ such that

$$||\boldsymbol{d}_k(\boldsymbol{x}_j)|| < \delta, \quad \forall k, j, \tag{5.40}$$

when $\hat{\gamma} < 2/\lambda$. According to (5.39) the voxel displacement $\boldsymbol{d}_k(\boldsymbol{x}_j)$ has a normal distribution, independent of k, with variance depending on $\hat{\gamma}$. The criterion given in (5.40) must therefore be weakened to, for example:

$$\Pr\left(||\boldsymbol{d}_k(\boldsymbol{x}_j)|| > \delta\right) < \rho, \quad \forall j, \tag{5.41}$$

with ρ some small value, say 0.05. We approximate (5.41) by:

$$\mathrm{E}||\boldsymbol{d}_k(\boldsymbol{x}_j)||^2 + 2\sqrt{\mathrm{Var}||\boldsymbol{d}_k(\boldsymbol{x}_j)||^2} < \delta^2, \quad \forall j. \tag{5.42}$$

For the expectation and variance the following expressions can be derived using (5.39):

$$\mathrm{E}||\boldsymbol{d}_k(\boldsymbol{x}_j)||^2 = \hat{\gamma}^2\sigma_1^2 \mathrm{tr}\left(\boldsymbol{J}_j\boldsymbol{C}\boldsymbol{J}_j'\right), \tag{5.43}$$

$$\mathrm{Var}||\boldsymbol{d}_k(\boldsymbol{x}_j)||^2 = 2\hat{\gamma}^4\sigma_1^4||\boldsymbol{J}_j\boldsymbol{C}\boldsymbol{J}_j'||_F^2, \tag{5.44}$$

with $||\cdot||_F$ denoting the Frobenius norm. Substitution in (5.42) gives:

$$\hat{\gamma}^2 < \min_{\boldsymbol{x}_j \in \Omega_F^V} \frac{\delta^2/\sigma_1^2}{\mathrm{tr}\left(\boldsymbol{J}_j\boldsymbol{C}\boldsymbol{J}_j'\right) + 2\sqrt{2}||\boldsymbol{J}_j\boldsymbol{C}\boldsymbol{J}_j'||_F}. \tag{5.45}$$

Setting the right-hand side equal to $(2/\lambda)^2$ results in the desired estimate of λ. The maximum value of a for a deterministic gradient descent method can then be computed using: $\gamma(0) = a/A = 2/\lambda$. We denote this maximum by a_{MAX}:

$$a_{\mathrm{MAX}} \equiv \frac{2A}{\lambda} \tag{5.46}$$

$$= \frac{A\delta}{\sigma_1} \min_{\boldsymbol{x}_j \in \Omega_F^V} \left[\mathrm{tr}\left(\boldsymbol{J}_j\boldsymbol{C}\boldsymbol{J}_j'\right) + 2\sqrt{2}||\boldsymbol{J}_j\boldsymbol{C}\boldsymbol{J}_j'||_F\right]^{-\frac{1}{2}}. \tag{5.47}$$

The second assumption that imposes a constraint on a is Assumption B3.2. Using $\gamma(0) = a/A$, $\mathrm{tr}(\boldsymbol{\Sigma}) = \mathrm{tr}(\sigma_3^2\boldsymbol{C}) = \mathrm{E}||\boldsymbol{\varepsilon}_k||^2$, and the definition of a_{MAX} in (5.46), we rewrite (5.12) as:

$$a \leq \frac{2A}{\lambda}\frac{||\boldsymbol{g}(\boldsymbol{\mu})||^2 - \beta_0}{||\boldsymbol{g}(\boldsymbol{\mu})||^2 + \mathrm{E}||\boldsymbol{\varepsilon}_k||^2} = a_{\mathrm{MAX}}\frac{||\boldsymbol{g}(\boldsymbol{\mu})||^2 - \beta_0}{||\boldsymbol{g}(\boldsymbol{\mu})||^2 + \mathrm{E}||\boldsymbol{\varepsilon}_k||^2}. \tag{5.48}$$

When the expected approximation error $\mathrm{E}||\boldsymbol{\varepsilon}_k||^2$ goes to zero, and $\beta_0 \downarrow 0$, this condition equals $a < a_{\mathrm{MAX}}$. The condition corresponds to the intuition that a lower gain should be used when the approximation error increases. Exact verification of (5.48) for all $||\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}|| \geq R$, as Assumption B3.2 demands, seems not feasible. We therefore propose to use the following estimate of a:

$$a = a_{\mathrm{MAX}} \frac{\mathrm{E}||\boldsymbol{g}||^2}{\mathrm{E}||\boldsymbol{g}||^2 + \mathrm{E}||\boldsymbol{\varepsilon}_k||^2} \equiv a_{\mathrm{MAX}}\eta, \tag{5.49}$$

with $0 < \eta \leq 1$. For $\mathrm{E}||\boldsymbol{g}||^2$ and $\mathrm{E}||\boldsymbol{\varepsilon}_k||^2$ their empirical estimates can be used directly, see the left-hand side of (5.34).

Summarising, we have replaced the original parameter a by a new user-defined parameter, δ. Unlike a, the new parameter δ has a unit (mm), and an intuitive meaning. In Sec. 5.3, the sensitivity of ASGD to the values of δ and A is experimentally investigated.

As announced in Sec. 5.2.3, we also use δ to select the value of σ_4, which occurs in (5.35). The voxel displacement caused by the parameter change from $\boldsymbol{\mu}_0$ to $\boldsymbol{\mu}_n$ is considered:

$$\boldsymbol{d}_{0,n}(\boldsymbol{x}_j) \equiv \boldsymbol{T}(\boldsymbol{x}_j, \boldsymbol{\mu}_n) - \boldsymbol{T}(\boldsymbol{x}_j, \boldsymbol{\mu}_0). \tag{5.50}$$

Following a similar approach as in Sec. 5.2.3, the distribution of $\boldsymbol{d}_{0,n}(\boldsymbol{x}_j)$ can be estimated, given the distribution of $\boldsymbol{\mu}_n$, which was defined in (5.35). The result is:

$$\boldsymbol{d}_{0,n}(\boldsymbol{x}_j) \sim \mathcal{N}\left(\boldsymbol{0}, \sigma_4^2 \boldsymbol{J}_j \boldsymbol{J}_j'\right). \tag{5.51}$$

With similar reasoning as earlier in this section, we select σ_4 such that:

$$\Pr\left(||\boldsymbol{d}_{0,n}(\boldsymbol{x}_j)|| > \delta\right) < \rho, \quad \forall j, \tag{5.52}$$

with ρ some small value, say 0.05. This condition is approximated by:

$$\mathrm{E}||\boldsymbol{d}_{0,n}(\boldsymbol{x}_j)||^2 + 2\sqrt{\mathrm{Var}||\boldsymbol{d}_{0,n}(\boldsymbol{x}_j)||^2} < \delta^2, \quad \forall j, \tag{5.53}$$

which, using (5.51), gives the following solution for σ_4^2:

$$\sigma_4^2 = \min_{\boldsymbol{x}_j \in \Omega_F^V} \frac{\delta^2}{||\boldsymbol{J}_j||_F^2 + 2\sqrt{2}||\boldsymbol{J}_j \boldsymbol{J}_j'||_F}. \tag{5.54}$$

5.2.5 Selection of sigmoid parameters

The selection of the sigmoid function parameters f_{MAX} and f_{MIN} is based on the condition for asymptotic normality: Assumption B5.2. This assumption constrains the value of E_0, which is directly related to f_{MAX} and f_{MIN}, according to Assumption B4.2. The third parameter ω, which defines the scale of the sigmoid, is chosen as a small fraction of the standard deviation of $\boldsymbol{\varepsilon}_k' \boldsymbol{\varepsilon}_{k-1}$, such that $E_0 \approx \frac{1}{2}(f_{\mathrm{MAX}} + f_{\mathrm{MIN}})$.

Assumption B5.2 states that matrix $\boldsymbol{W} = \frac{1}{2}\boldsymbol{I} - \frac{a}{E_0}\boldsymbol{H}(\hat{\boldsymbol{\mu}})$ is assumed to have negative eigenvalues only. Let $\lambda^* > 0$ denote the minimum positive eigenvalue of $\boldsymbol{H}(\hat{\boldsymbol{\mu}})$ (Assumption B5.2 can never be satisfied with negative eigenvalues of $\boldsymbol{H}(\hat{\boldsymbol{\mu}})$). The condition then becomes:

$$E_0 < 2a\lambda^*. \tag{5.55}$$

Combining (5.55), (5.49), and (5.46) gives:

$$E_0 < 4A\frac{\lambda^*}{\lambda}\eta. \tag{5.56}$$

So, given fixed A and cost function properties λ and λ^*, the maximum allowed value of E_0 is directly proportional to the ratio η. Following Assumption B4.2 and assuming that ω is small, we have $E_0 \approx \frac{1}{2}(f_{\mathrm{MAX}} + f_{\mathrm{MIN}})$. Substitution in (5.56) gives:

$$f_{\mathrm{MIN}} < 8A\frac{\lambda^*}{\lambda}\eta - f_{\mathrm{MAX}}. \tag{5.57}$$

A reasonable choice for the maximum of the sigmoid function is $f_{\mathrm{MAX}} = 1$. This implies that the forward time step $t_{k+1} - t_k$ equals at most the time step made by the RM method. Demanding $-f_{\mathrm{MAX}} < f_{\mathrm{MIN}} < 0$, we propose:

$$f_{\mathrm{MIN}} = \eta - f_{\mathrm{MAX}} = \eta - 1, \tag{5.58}$$

where we assumed that A can be chosen such that $8A\lambda^*/\lambda > 1$, in order to satisfy (5.57). If A is chosen too low, the consequence is that asymptotic normality can not be guaranteed anymore. By choosing A very high this risk is avoided, but one has to keep in mind that the property of asymptotic normality is not always relevant in practice. In practical applications, the number of iterations K, see (5.5), is finite due to limited available computation time. Choosing $A \to \infty$ would result in a nearly constant gain sequence $\gamma(t_k)$ for all iterations $0 \leq k \leq K$. The adaptive behaviour of ASGD would, consequently, be eliminated completely. In Sec. 5.3, the sensitivity of ASGD to the value of A is experimentally investigated.

For the selection of f_{MIN} and f_{MAX} we assumed that $E_0 \equiv \mathrm{E}f(\boldsymbol{\varepsilon}_k'\boldsymbol{\varepsilon}_{k-1}) \approx \frac{1}{2}(f_{\mathrm{MAX}} + f_{\mathrm{MIN}})$. The approximation only holds if ω is much smaller than $|\boldsymbol{\varepsilon}_k'\boldsymbol{\varepsilon}_{k-1}|$, with high probability. According to (5.33), $\boldsymbol{\varepsilon}_k$ and $\boldsymbol{\varepsilon}_{k-1}$ are independent normally distributed variables with mean $\mathbf{0}$ and variance $\sigma_3^2\boldsymbol{C}$. The expected value of the inner product $\boldsymbol{\varepsilon}_k'\boldsymbol{\varepsilon}_{k-1}$ is zero. We propose to choose ω as a small fraction ζ of the standard deviation of $\boldsymbol{\varepsilon}_k'\boldsymbol{\varepsilon}_{k-1}$:

$$\omega = \zeta\sqrt{\mathrm{Var}\left(\boldsymbol{\varepsilon}_k'\boldsymbol{\varepsilon}_{k-1}\right)}, \tag{5.59}$$

with $\zeta \approx \frac{1}{10}$ for example. For the variance it can be shown that:

$$\mathrm{Var}\left(\boldsymbol{\varepsilon}_k'\boldsymbol{\varepsilon}_{k-1}\right) = \sigma_3^4||\boldsymbol{C}||_F^2. \tag{5.60}$$

An alternative strategy might be to actually set $\omega \downarrow 0$ (as small as machine precision allows), but this would start to interfere with Assumption B5.3.

5.2.6 Overview of the algorithm

The following steps describe the entire algorithm:

1. Compute $\boldsymbol{C}$ using (5.25).
2. Compute σ_4 using (5.54).
3. Generate N instances of $\boldsymbol{\mu}_n$ according to (5.35). Compute for each $\boldsymbol{\mu}_n$ the exact cost function derivative $\boldsymbol{g}$, the approximated derivative $\widetilde{\boldsymbol{g}}_k$, and the approximation error $\boldsymbol{\varepsilon}_k = \boldsymbol{g} - \widetilde{\boldsymbol{g}}_k$. Note that, to compute $\widetilde{\boldsymbol{g}}_k$, a new set of voxels S_k must be selected for each $\boldsymbol{\mu}_n$.
4. Compute σ_1 using (5.34). Compute σ_3 similarly.
5. Compute a_{MAX} using (5.47).
6. Compute η and a using (5.49).
7. Set $f_{\text{MAX}} = 1$ and compute f_{MIN} using (5.58).
8. Compute ω using (5.59) and (5.60).
9. Start the optimisation defined by (5.5), (5.6), (5.9), and (5.10). Convergence is assumed after K iterations: $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}_K$.

Steps 1-8 serve to estimate a, f_{MAX}, f_{MIN}, and ω. Note that this has to be done only once, before starting the actual optimisation routine in step 9. The required user settings are t_0, t_1, K, δ, A, N, and ζ. The initial conditions t_0 and t_1 will probably have a minor influence on the performance as long as they are chosen much smaller than the number of iterations: $t_0, t_1 \ll K$. The meaning of δ is explained in Sec. 5.2.4 and the influence of A is discussed in Sec. 5.2.5. For N, a value ≈ 10 is suggested in Sec. 5.2.3. For ζ, a value $\approx \frac{1}{10}$ is recommended in Sec. 5.2.5.

5.3 Experiments and results

5.3.1 Experiment setup

The ASGD method has been evaluated on three medical image registration problems. Table 5.1 gives an overview of the data sets that were used, the type of registration experiments, and the evaluation measures for quantifying registration accuracy. The bottom row indicates in which subsection the experiments are described.

The ASGD method was implemented as a part of the `elastix` package (see Chapter 2). Rigid, affine, and nonrigid B-spline transformation models were tested. A three-level multiresolution framework was used in all experiments. The images were smoothed with a Gaussian filter with a standard deviation of 2, 1, and 0.5 (voxel units), in each level respectively. For the B-spline transform, the B-spline control point grid spacing was halved in each resolution level, such that in the final resolution the grid spacing reported in Table 5.1 was reached. Four similarity measures were used: mean squared intensity difference (MSD), normalised correlation (NC), mutual information (MI), and normalised mutual information (NMI). Both MI and NMI were implemented using cubic B-spline Parzen windows, as in

Table 5.1: *Overview of data sets and experiments.*

anatomy	brain	prostate	right lung
modality	CT and 1.5T MR T1	3T MR SSFP	CT
dimensions	CT: 512×512×50 MR: 256×256×50	200×200×70	120×160×200
voxel size [mm]	CT: 0.45×0.45×3 MR: 0.85×0.85×3	0.5×0.5×1	2×2×2
nr. of patients	9	6 (2 scans/person)	5 (2 scans/person)
registration	CT with MR	day 1 with day 2	day 1 with day 2
similarity measure	MI	MI	MSD, NC, MI, NMI
transformation	rigid	B-spline	affine, B-spline
nr. of parameters P	6	2000	12, 4000
B-spline control point grid spacing [mm]	-	16×16×16	40×40×40
evaluation measure	MSE	DSC	DSC
Section	5.3.2	5.3.2	5.3.3 and 5.3.4

[85], with a 32×32 joint histogram. For the rigid registrations, the transformation was parameterised using the translation vector $\boldsymbol{t} = (t_1, t_2, t_3)'$ and the Euler angles $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)'$. Since the Euler angles can have an entirely different range than the translations, we used the following re-parametrisation:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{S} \end{bmatrix} \begin{bmatrix} \boldsymbol{t} \\ \boldsymbol{\theta} \end{bmatrix}, \tag{5.61}$$

with $\boldsymbol{S}$ a diagonal scaling matrix, with on the diagonal:

$$s_{ii} = \left(\int_{\Omega_F} \left|\left| \frac{\partial \boldsymbol{T}}{\partial \theta_i} (\boldsymbol{x}, \boldsymbol{\mu}_0) \right|\right|^2 d\boldsymbol{x} \Big/ \int_{\Omega_F} d\boldsymbol{x} \right)^{-\frac{1}{2}}. \tag{5.62}$$

The rotation parameters are thus scaled by the average voxel displacement caused by a small perturbation of the rotation angle. In case of an affine transformation we used the same strategy for the matrix elements. In case of a B-spline transformation the control point coefficients directly formed the parameters $\boldsymbol{\mu}$.

For the brain images, the ground truth CT-MR registrations were available. The scans were acquired using a stereotactic frame, which was later erased from the images by post-processing, in the context of the "Retrospective Image Registration Evaluation" project [95]. In our experiments we quantified the registration accuracy by computing the mean square error of the transformation at the eight corner points of the image:

$$\mathrm{MSE} \equiv \tfrac{1}{8} \sum_{c=1}^{8} ||\boldsymbol{T}(\boldsymbol{x}_c, \hat{\boldsymbol{\mu}}) - \boldsymbol{T}(\boldsymbol{x}_c, \hat{\boldsymbol{\mu}}^G)||, \tag{5.63}$$

with $\hat{\boldsymbol{\mu}}^G$ the ground truth.

For the MR prostate scans, expert manual segmentations of the prostate were available. The Dice similarity coefficient (DSC) [18] of the segmentation S_F of the fixed image and the segmentation S_M of the deformed moving image was used for evaluation:

$$\mathrm{DSC} \equiv \frac{2|S_F \cap S_M|}{|S_F| + |S_M|}. \tag{5.64}$$

The DSC measures overlap of the two segmentations and thus gives an indication of the registration quality. A value of 1 means perfect registration. A value of 0 means that the segmentations have no overlap at all.

For the CT lung images, we used the DSC of the lung airways as an evaluation measure. The lung airways were segmented using an automatic region-growing algorithm, described in [27, 73].

In all experiments we used the initial conditions $t_0 = t_1 = 0$. As suggested in Sec. 5.2, we used $N = 10$ and $\zeta = \frac{1}{10}$. The number of voxels used to compute $\widetilde{g}_k$, denoted by $|S_k|$ in Eq. (5.26), was set to 2000, as in Chapter 3. For the remaining free parameters, δ, A, and K, the settings are reported in the following subsections. In all experiments, the extra computation time required by ASGD to perform steps 1-8, see Sec. 5.2.6, was comparable to the time spent in step 9.

In Sec. 5.3.2, the ASGD method is compared with the standard RM method. The brain and prostate data are used for this purpose. In Sec. 5.3.3, the lung images are used to test ASGD with different similarity metrics. In Sec. 5.3.4, the relation between δ and the maximum voxel displacement is verified.

5.3.2 Adaptive vs. non-adaptive

In this subsection, we test the effect of the step size adaptation. The ASGD method is compared to the standard RM method in a series of experiments on the brain and prostate data, for a range of values of δ, A, and K.

The RM method, see (5.2), requires definition of the step size sequence $\{\gamma_k\}$. For fair comparison with ASGD, we use the following function:

$$\gamma_k = a/(E_0 k + A), \tag{5.65}$$

with a, A, and E_0 as computed for the ASGD method. With this choice γ_0 equals $\gamma(t_0)$, so RM and ASGD start with the same step size. Also, it can be shown [13] that γ_k and $\gamma(t_k)$ converge to the same value as $k \to \infty$. For this it is necessary to see that, with ASGD, E_0 equals the expected value of the time increment $t_{k+1} - t_k$ when $\boldsymbol{g}(\boldsymbol{\mu}_k) \approx \mathbf{0}$.

The registration experiments were performed for all possible combinations of $\delta \in \{0.03125, 0.0625, \ldots, 64\}$ (in mm), $A \in \{1.25, 2.5, \ldots, 320\}$, and $K \in \{250, 2000\}$. The brain images were registered using a rigid transformation model. For each (δ, A, K) combination the mean MSE over the 9 CT-MR registrations was calculated. The prostate scans were registered using a nonrigid B-spline transformation. After registration, the mean DSC over the 6 image pairs was computed. The measured

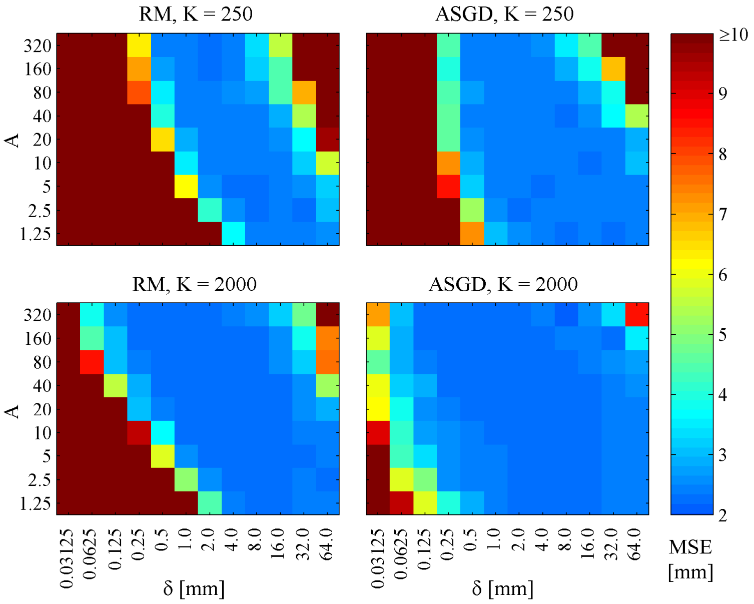


Figure 5.2: *RM vs. ASGD for rigid registration of brain scans. A low MSE indicates better registration.*

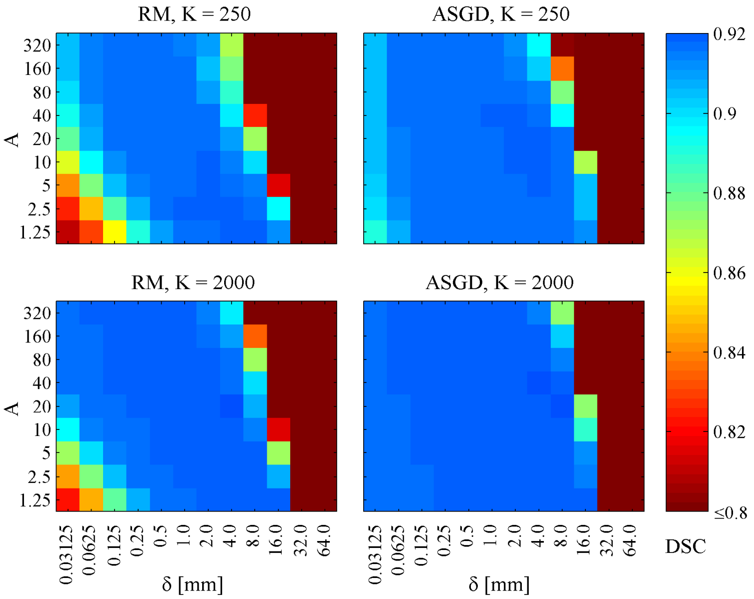


Figure 5.3: *RM vs. ASGD for nonrigid registration of prostate scans. A high DSC indicates better registration.*

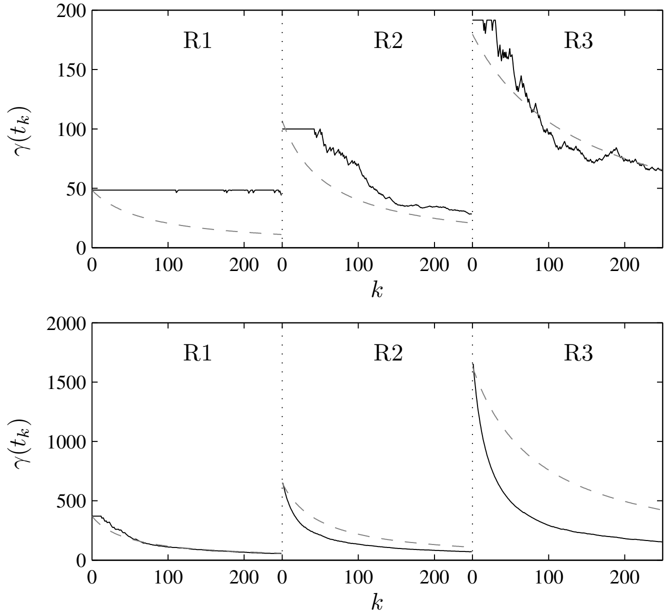


Figure 5.4: *Example of step size adaptation by ASGD. The solid black line is for ASGD; the dashed grey line for RM. The upper graph shows the result for $\delta = 0.25$ mm. The lower graph was created using $\delta = 2$ mm. Note that the vertical axes have different scales.*

computation time per registration on an AMD Opteron 2600 MHz was approximately 5 min.

In Figs. 5.2 and 5.3 the results are visualised on a colour scale. Each pixel represents the mean MSE or DSC for a combination of δ and A. The adaptive step size mechanism clearly improved the robustness with respect to the user-defined parameters A and δ. Increasing the number of iterations from $K = 250$ to $K = 2000$ improved the robustness of both RM and ASGD. However, Fig. 5.3 shows that the ASGD method with $K = 250$ gave better results still than RM with $K = 2000$.

As an illustration of the step size adaptation by ASGD we plotted the values of $\gamma(t_k)$ during registration of one of the prostate image pairs. Fig. 5.4 shows the result for $\delta = 0.25$ mm (upper graph) and $\delta = 2$ mm (lower graph), both with $A = 20$ and $K = 250$. The labels R1-3 represent the three resolution levels. The solid black line is for ASGD. The dashed grey line shows the predefined step size function that was used for RM, as given by (5.65). The adaptive step size mechanism of ASGD is clearly observed in each resolution: when the algorithm starts with a small step size ($\delta = 0.25$ mm), the step size decays less fast than with a large initial step size ($\delta = 2$ mm). For example, in resolution R2, with $\delta = 0.25$, the ASGD step size remains nearly constant in the first 50 iterations, whereas with $\delta = 2.0$ the step size immediately starts decaying at $k = 0$.

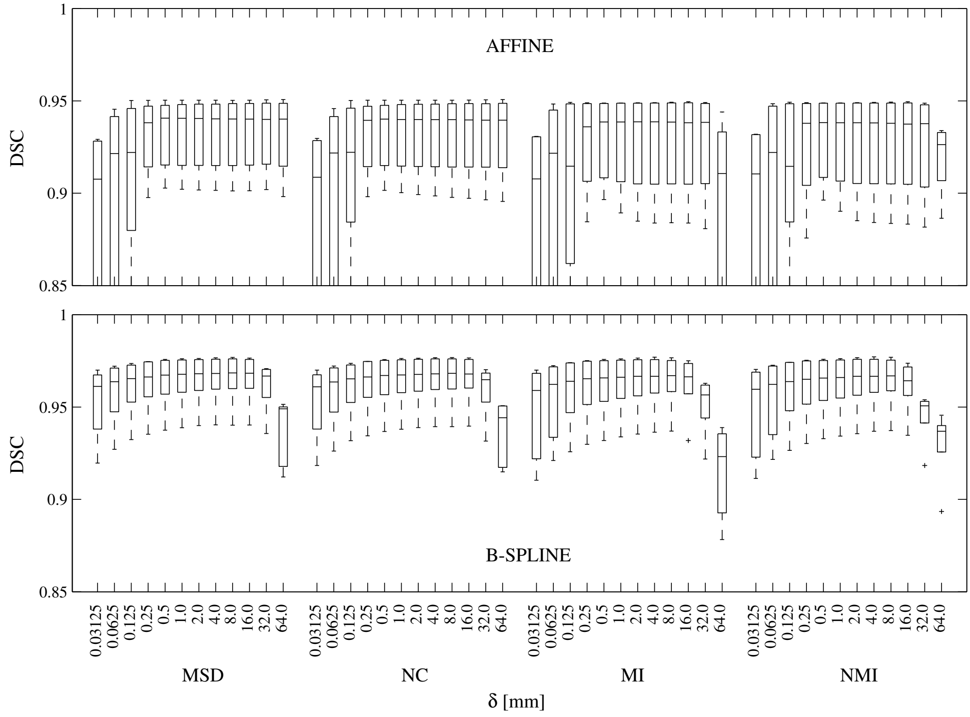


Figure 5.5: *ASGD with different similarity measures for registration of CT lung scans. The upper plot shows the results for an affine transformation. The lower plot shows the results for a B-spline transformation.*

5.3.3 ASGD with different similarity measures

In this subsection, we investigate the influence of the similarity measure on the choice of δ. Registration experiments were performed on the CT lung data using different similarity measures, for a range of δ values.

Four similarity measures were tested: MSD, NC, MI, and NMI. For δ the range $\{0.03125, 0.0625, \ldots, 64\}$ (in mm) was used. The entire experiment was done using both an affine and a B-spline transformation. All registrations were done with $A = 20$, which gave good performance in the previous section. A relatively low number of iterations was used, $K = 250$, such that the effect of varying δ becomes more apparent.

The results are summarised in Fig. 5.5. Each boxplot summarises the distribution of DSC values after registration of the five image pairs. The upper graph shows the results using the affine transformation. The lower graph shows the results obtained with the B-spline transformation. Both for the affine and the B-spline registrations, the optimal value of δ was independent of the choice of the similarity

Table 5.2: *Average values of a in the finest resolution level of the lung image registrations, using $\delta = 1$ mm.*

	MSD	NC	MI	NMI
affine	0.0017	620	240	780
B-spline	0.73	270000	43000	140000

measure. For affine registration, the range $0.5 \leq \delta \leq 32$ mm gave the best results. With the B-spline transformation, the range $1 \leq \delta \leq 16$ mm gave the best results. For $\delta = 1$ mm, the calculated values of a in the finest resolution level are reported in Table 5.2, averaged over the five image pairs. The large differences between the values show that choosing a manually would not have been a trivial task.

5.3.4 Maximum voxel displacement

In Sec. 5.2.4, δ was introduced as a user setting with an intuitive meaning, being the maximum voxel displacement per iteration of the deterministic gradient descent process $\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - \hat{\gamma}\boldsymbol{g}(\boldsymbol{\mu}_k)$, with constant step size $\hat{\gamma} = a_{\text{MAX}}/A$. The estimate of a_{MAX} relies on some simplifying assumptions and approximations, most notably Assumption A2. The following experiment serves to verify whether the voxel displacements indeed remain below δ.

The CT lung registrations were repeated with the deterministic gradient descent scheme mentioned above, using $K = 100$, and MI as a similarity measure. The voxel displacements $||\boldsymbol{d}_k(\boldsymbol{x}_j)||$ were computed for all $\boldsymbol{x}_j \in \Omega_F^V$, in each iteration k. Table 5.3 reports the 95% quantiles of the ratio $||\boldsymbol{d}_k(\boldsymbol{x}_j)||/\delta$ for each resolution level separately. Each entry in the table is based on 5 image pairs. Entries with '-' indicate that for at least one of the image pairs the registration failed completely, i.e. the overlap between the fixed and moving image became too small to continue registration (due to very large step sizes). The table shows that with the affine transformation the ratio was close to 1, meaning that most voxel displacements indeed remained below δ. With the B-spline transformation, for $\delta \geq 0.5$ the actual displacements exceeded δ with a factor 2 on average. For $\delta < 0.5$, the actual displacements remained below δ.

5.4 Discussion

The experiments show that ASGD works for a rather broad range of δ and A. The results in Sec. 5.3.2 indicate that $A = 20$ works well in general, both for rigid and nonrigid registration. With that setting, for the applications we considered, the optimum value of δ was approximately equal to the size of a voxel. Of course, that relation is not always exactly satisfied, since simply upsampling the images will not lower the optimum value of δ. However, the experiments in Secs. 5.3.2 and 5.3.3 show that the registration results are relatively insensitive to the value of δ,

Table 5.3: *The 95% quantiles of the ratio* $||\boldsymbol{d}_k(\boldsymbol{x}_j)||/\delta$. *A value close to 1 is desirable. Each entry in the table is based on 5 image pairs,* $K = 100$ *iterations, and all voxels* $\boldsymbol{x}_j \in \Omega_F^V$.

		δ [mm]											
transform	resolution	0.03125	0.0625	0.125	0.25	0.5	1.0	2.0	4.0	8.0	16.0	32.0	64.0
affine	1	0.8	0.8	0.8	0.7	0.7	0.6	0.6	1.7	1.8	1.4	1.3	1.0
affine	2	0.7	0.7	0.7	0.6	0.7	0.9	1.0	0.6	0.8	0.9	1.3	1.1
affine	3	0.7	0.7	0.7	0.7	0.6	1.2	0.4	0.8	0.8	0.8	1.1	1.2
B-spline	1	0.5	0.4	0.4	0.4	0.3	1.8	2.3	1.9	1.6	1.4	1.4	1.2
B-spline	2	0.7	0.6	0.5	0.4	1.7	4.2	2.3	1.6	1.6	1.6	1.4	-
B-spline	3	0.9	0.8	0.6	0.4	4.3	2.6	2.2	1.7	1.6	1.4	1.4	-

as long as $\frac{1}{4}V \leq \delta \leq 4V$, with V the (average) voxel size. For rigid and affine registration somewhat higher δ values tend to work better than for nonrigid, which corresponds to intuition.

The results in Sec. 5.3.4, Table 5.3, show that the actually realised voxel displacements were not in all cases lower than δ. This is due to the simplifying assumptions used to estimate a_{MAX} given δ. Especially Assumption A2 may not be satisfied. While the estimate of a_{MAX} still appears to work quite well in practice, further improvements may be obtained by improving the estimate of the distribution of $\boldsymbol{g}$. In our approach, the estimated distribution of $\boldsymbol{g}$ is for a large part based on the model for the covariance matrix, given by (5.25). This allows us to use a low number N of gradient evaluations, since only one parameter (σ_1) has to be determined, see Eq. (5.31) and below. Another approach would be to use the common maximum likelihood estimate of the covariance matrix: $\boldsymbol{C} = \frac{1}{N}\sum_n \boldsymbol{g}(\boldsymbol{\mu}_n)\boldsymbol{g}(\boldsymbol{\mu}_n)'$. However, this would require a larger N. An interesting technique that combines the two approaches is the *shrinkage* method described in [70]. In that article, a linear combination of the model based estimate of $\boldsymbol{C}$ and the maximum likelihood estimate is employed, with the weighting determined by explicitly minimising the expectation of a squared error loss function.

In all experiments described in this chapter, the initial conditions t_0 and t_1 were simply set to 0. It might be beneficial to try larger values, such as $t_0 = t_1 = A$. In this way, the method could become more robust to large values of δ.

5.5 Conclusion

An optimisation method with adaptive step size prediction for image registration has been presented: adaptive stochastic gradient descent (ASGD). The method is designed to work with stochastic approximations of the cost function derivatives, and, thus, requires little computation time per iteration. In comparison with a standard Robbins-Monro (RM) stochastic gradient descent scheme, the ASGD method is more robust, because of its adaptive step size prediction. Our main contribution is the selection mechanism for the method's free parameters. The selection mechanism takes into account the choice of similarity measure, the transformation model, and the image content, in order to estimate proper values for the most important settings. The influences of the remaining free parameters δ, A, and K were experimentally investigated. The experiments showed that ASGD works for a broad range of δ and A. The optimum value of δ appeared to be unaffected by the choice of the similarity measure. In general, a reasonable setting is to use $A = 20$ and δ equal to the average voxel size of the images. Increasing the number of iterations from $K = 250$ to $K = 2000$ improved the robustness of both RM and ASGD, with respect to the choice of δ and A. However, the ASGD method with $K = 250$ gave already better results than RM with $K = 2000$.

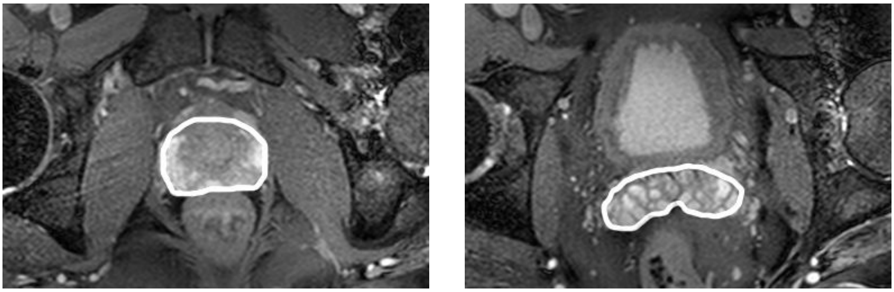
In Chapter 3, it was shown for a number of medical image registration applications that RM outperforms several well-known deterministic optimisation methods, such as quasi-Newton and nonlinear conjugate gradient. It was pointed out that the main disadvantage of RM is the need for a predetermined step size function. The ASGD method presented in this article provides a solution for that issue.

Chapter 6

Segmentation of the Prostate by Atlas Matching

S. Klein, U.A. van der Heide, I.M. Lips, M. van Vulpen, M. Staring, J.P.W. Pluim. Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information. *Medical Physics*, 35(4):1407–1417, 2008.

An automatic method for delineating the prostate (including the seminal vesicles) in 3D magnetic resonance (MR) scans is presented in this chapter. The method is based on nonrigid registration of a set of prelabelled atlas images. Each atlas image is nonrigidly registered with the target patient image. Subsequently, the deformed atlas label images are fused to yield a single segmentation of the patient image. The proposed method is evaluated on 50 clinical scans, which were manually segmented by three experts. The Dice similarity coefficient (DSC) is used to quantify the overlap between the automatic and manual segmentations. We investigate the impact of several factors on the performance of the segmentation method. For the registration, two similarity measures are compared: mutual information and a localised version of mutual information. The latter turns out to be superior (median $\Delta DSC \approx 0.02$, $p < 0.01$ with a paired two-sided Wilcoxon test) and comes at no added computational cost, thanks to the use of a novel stochastic optimisation scheme. For the atlas fusion step we consider a majority voting rule and the "simultaneous truth and performance level estimation" (STAPLE) algorithm, both with and without a preceding atlas selection stage. The differences between the various fusion methods appear to be small and mostly not statistically significant ($p > 0.05$). To assess the influence of the atlas composition, two atlas sets are compared. The first set consists of 38 scans of healthy volunteers. The second set is constructed by a leave-one-out approach using the 50 clinical scans that are used for evaluation. The second atlas set gives substantially better performance ($\Delta DSC = 0.04$, $p < 0.01$), stressing the importance of a careful atlas definition. With the best settings, a median DSC of around 0.85 is achieved, which is close to the median interobserver DSC of 0.87. The segmentation quality is especially good at the prostate-rectum interface, where the segmentation error remains below 1 mm in 50% of the cases and below 1.5 mm in 75% of the cases.



(a) The peripheral zone and the central gland

(b) The seminal vesicles

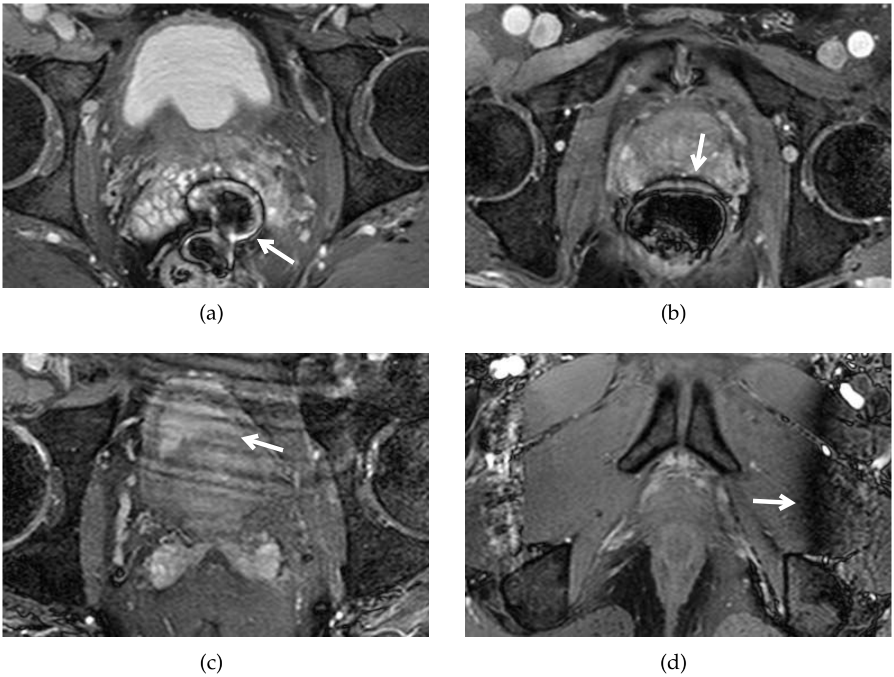
Figure 6.1: *Two example MR slices, zoomed in on the region of interest, with manually delineated prostate (white line).*

6.1 Introduction

Prostate cancer treatment by radiation therapy requires an accurate localisation of the prostate: neighbouring tissue (rectum and bladder) should be spared, while the tumour should receive a prescribed dose. For the treatment planning, computed tomography (CT) images are primarily used, but increasingly magnetic resonance (MR) images are added, because of their soft-tissue contrast [90, 91]. Several studies [61, 91] have demonstrated that the additional use of MR images for prostate delineation leads to a reduced interobserver variation and a smaller estimated prostate volume. In current practice at our hospital a manual delineation of the prostate is made, based on the CT and MR scans, which is a labour-intensive task and requires training. Therefore, automating this process is desired.

Figure 6.1 shows two example MR slices together with their manual delineations. The MR protocol, a balanced steady-state free precession (bSSFP) sequence, was optimised for visibility of the prostate and rectum. An extensive review of the prostate's anatomy visible on MR images can be found in [90]. The major components of the prostate are central gland, the peripheral zone, and the seminal vesicles, each having different appearances on the bSSFP MR scans. The shape and size of the seminal vesicles vary heavily among people. The central gland and the peripheral zone together have the size of a walnut (around 25 ml) in healthy subjects. Prostate cancer develops most frequently in men over fifty. With increasing age, a large group of men also suffers from benign prostate hypertrophy (BPH), which can result in substantial growth of the central gland.

Challenges for automatic segmentation of the prostate in MR images include the presence of imaging artefacts due to air in the rectum and inhomogeneities of the magnetic field, the large anatomical variability between subjects, the differences in rectum and bladder filling, and the lack of a normalised 'Hounsfield' unit for MR. Four examples of imaging artefacts, taken from the clinical test data described in Sec. 6.3.1.2, are shown in Fig. 6.2.



(a)

(b)

(c)

(d)

Figure 6.2: *Four examples of imaging artefacts, marked by white arrows; a) and b) susceptibility artefacts due to air in the rectum, which manifest themselves as black lines not corresponding to tissue boundaries, c) low-contrast prostate-bladder boundary combined with a streaking artefact, d) large intensity inhomogeneity.*

Recent surveys of the literature on automatic segmentation of the prostate can be found in [54, 97]. Existing work has mainly focussed on statistical model based approaches. In [98], a pseudo-3D active shape model is used to segment the prostate without seminal vesicles in MR images. In [54], a method is proposed that combines a statistical model for the prostate with region-growing methods for the rectum and the bladder. The seminal vesicles are not included and manual initialisation is required. In [22] a method to segment pelvic CT images is presented that uses *intrasubject* nonrigid registration of a manually segmented planning scan.

We propose a fully automatic method to delineate the prostate including the seminal vesicles in 3D MR scans. The method is based on *intersubject* registration of atlas images. The atlas consists of a set of manually labelled MR images from multiple individuals. Using a nonrigid registration algorithm, all atlas images are matched to the patient's MR image that is to be segmented. The deformed manual segmentations of the atlas images are combined into a single segmentation of the patient's image (label fusion). Multiple atlas images are used, instead of a single image, to account for the large anatomical variability between subjects and for the differences in bladder and rectum filling. Multi-atlas based segmentation methods have given promising results in other applications [64].

Three factors that may influence the performance are investigated: the similarity measure used by the registration process, the atlas label fusion method, and the composition of the atlas set. In Sec. 6.2.1, the two similarity measures are described: mutual information and a localised version of mutual information. For the latter we introduce a novel stochastic optimisation method. The atlas label fusion methods are treated in Sec. 6.2.2. The atlas sets that are used in the experiments are described in Sec. 6.3.

The proposed method is evaluated on 50 clinical scans. To determine the ground truth, each scan was manually segmented by three human experts. The Dice similarity coefficient (Sec. 6.3.3) is used to quantify the overlap between automatic and manual segmentations. The spatial distribution of the segmentation errors is visualised using a spherical coordinate mapping of the prostate boundary. Section 6.4 presents the results of the experiments. First, the impact of the three factors mentioned above is explored. Subsequently, the accuracy of the automatic segmentation obtained with the optimum settings is compared to the interobserver variability. Recommendations for future work are given in Sec. 6.5 and the chapter is concluded in Sec. 6.6.

6.2 Method

The patient's image to be segmented is denoted by $P(\boldsymbol{x})$. The goal of the automatic segmentation method is to produce a binary label image $L(\boldsymbol{x})$ that accurately defines the prostate of the patient.

The proposed segmentation method follows the general scheme of multi-atlas based segmentation methods, see for example [64]. A set of M accurately labelled images, which serve as an atlas, is assumed to be available. The ith image in this atlas set is referred to as $A_i(\boldsymbol{x})$. The corresponding label image, created by a human expert, is called $L_i(\boldsymbol{x})$. The segmentation method consists of two stages: 1) registration and 2) label image fusion. In the registration stage, each atlas image A_i is matched to the patient image P, using a nonrigid registration algorithm. The resulting coordinate transformations are applied to the label image L_i. In the label image fusion stage, the deformed label images are combined into a single segmentation L of the target patient image. Note that in all steps the images are treated as 3D volumes, rather than processing them on a 2D slice-by-slice basis.

6.2.1 Registration

In the registration stage, each atlas image A_i is matched to the patient image P. A coordinate transformation $T_i(\boldsymbol{x})$ is estimated that maximises the similarity of P and the deformed atlas $A_i \circ T_i$.[†] The registration is performed in two steps. Firstly, rough alignment of the two images is achieved by a rigid registration. After that

[†] The symbol $\circ$ represents function composition: $(A_i \circ T_i)(\boldsymbol{x}) = A_i(T_i(\boldsymbol{x}))$.

a nonrigid registration is performed, using a coordinate transformation that is parameterised by cubic B-splines [68]. The parameters that describe the transformation are represented by the vector $\boldsymbol{\mu}$.

An important aspect of the registration method is the choice of the similarity measure. We compare two similarity measures: mutual information (MI) [41, 92] and localised mutual information (LMI) [24, 78]. The mutual information of two d-dimensional images $I(\boldsymbol{x}), J(\boldsymbol{x}) : \Omega \subset \mathbb{R}^d \mapsto \mathbb{R}$ is defined as follows:

$$\mathrm{MI}(I, J; \Omega) = \sum_k \sum_m p_{IJ}(k, m) \log \frac{p_{IJ}(k, m)}{p_I(k) p_J(m)}, \tag{6.1}$$

where p_I and p_J denote the discrete marginal intensity probabilities of I and J, respectively, and p_{IJ} represents the discrete joint intensity probability. The intensity probabilities are estimated from a discrete set of intensity pairs $(I(\boldsymbol{x}_i), J(\boldsymbol{x}_i))$, where the coordinates $\boldsymbol{x}_i$ are sampled from the continuous image domain Ω. A common choice is to use all voxel locations, or a uniformly sampled subset of those. An important assumption of MI is that the true intensity probabilities do not vary over Ω. This assumption is often violated in MR scans, due to the presence of magnetic field inhomogeneities. Therefore, it may be better to evaluate the mutual information on multiple subregions, each having a more stationary intensity distribution. Adding the resulting mutual information values of all subregions gives us the localised mutual information LMI [24, 78]:

$$\mathrm{LMI}(I, J; \Omega) = \frac{1}{N} \sum_{\boldsymbol{x}_j \in \Omega} \mathrm{MI}(I, J; \mathcal{N}(\boldsymbol{x}_j)). \tag{6.2}$$

In this equation $\mathcal{N}(\boldsymbol{x}_j) \subseteq \Omega$ represents a spatial neighbourhood centred on $\boldsymbol{x}_j$. The number of neighbourhoods is denoted by N. The neighbourhood centre coordinates $\boldsymbol{x}_j$ are samples from Ω. We may choose them to be all voxel locations, or some subset of those. The neighbourhoods $\mathcal{N}(\boldsymbol{x}_j)$ must be chosen large enough to allow for a reliable estimation of the intensity probabilities, but small enough to ensure that the influence of the inhomogeneities is negligible. We considered cubic regions of 25^3 mm, 50^3 mm, and 100^3 mm, and compared their performance in 36 registrations on a subset of the data described in Secs. 6.3.1.1 and 6.3.1.2. Six scans of the first data set were registered to six scans of the second data set. The best results in terms of the prostate overlap after registration, see Sec. 6.3.3, were obtained with the $50 \times 50 \times 50$ mm region. This setting is used in all experiments that are described in this chapter.

For maximisation of the similarity measure we employ an iterative optimisation routine, called stochastic gradient descent. The parameters $\boldsymbol{\mu}$ that describe the transformation are updated in each iteration k by taking a step in the direction of the derivative of the similarity measure with respect to $\boldsymbol{\mu}$. In Chapter 3 it was demonstrated for MI that convergence to the solution is still achieved when the derivative is approximated using only a very small number P of randomly sampled intensity pairs. Two important conditions for this are that new samples are selected in every iteration and that the step size a_k is a slowly decaying function of the iteration number k:

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - a_k \widetilde{\boldsymbol{g}}_k, \tag{6.3}$$

$$a_k = a/(k+A)^{\alpha}, \tag{6.4}$$

where $\widetilde{g}_k$ represents the approximated derivative of MI, and $a > 0$, $A \geq 1$, and $0 \leq \alpha \leq 1$ are user-defined constants. For LMI we can use the same strategy and even extend it by using a small set of neighbourhoods, randomly selected in every iteration. We use $N = 1$, in other words, LMI is implemented by computing MI using P intensity pairs, sampled from a $50 \times 50 \times 50$ mm neighbourhood that is randomly selected in every iteration. This approach results in equal computational costs per iteration for MI and LMI, provided that the same number of intensity pairs P are sampled to estimate p_{IJ}. The stochastic optimisation procedure that we use is an important difference to [24, 78].

The registration algorithm was integrated in `elastix`, a publicly available package for medical image registration (see Chapter 2. The mutual information MI is implemented according to [85], using a joint histogram size of 32×32 and cubic B-spline Parzen windows. The number of samples randomly selected in each iteration is set to $P = 2000$. A four-level multiresolution scheme is employed in both the rigid and the nonrigid registration step. Gaussian smoothing is applied to the image data using a standard deviation of 4, 2, 1, and 0.5 voxels in the four respective resolutions. The nonrigid registrations are performed using a B-spline control point spacing of 64, 32, 16, and 8 mm in all directions, for the four respective resolutions. Per resolution, 2000 iterations are performed. The step size sequence in Eq. (6.4) is defined by $a = 2000$, $A = 200$, and $\alpha = 0.6$. The above described settings were determined by trial-and-error experiments on two image pairs, randomly selected from the data set described in Sec. 6.3.1.1.

6.2.2 Label image fusion

The registration stage yields a set of transformations T_i, which can be applied to the atlas label images L_i, resulting in a set of deformed label images $L_i \circ T_i, i = 1, \ldots, M$. These must be combined into a single segmentation of the patient's image. For this purpose we consider majority voting (VOTE) and "simultaneous truth and performance level estimation" (STAPLE), explained in Secs. 6.2.2.2 and 6.2.2.3, respectively. Both methods can be combined with a preceding atlas selection stage, which is described in Sec. 6.2.2.1. In the experiments we compare VOTE and STAPLE, both with and without the atlas selection procedure.

The voxels of the atlas label images L_i take discrete values $c \in \mathcal{C}$, each corresponding to a certain tissue type (class), with $\mathcal{C}$ the set of classes. For example, '1' represents prostate tissue, '2' represents the bladder, and '0' everything else. Although the aim of our work is segmentation of only the prostate, the label fusion procedures VOTE and STAPLE may benefit from additionally labelled tissue types in the atlas. In the experiments this aspect is investigated.

6.2.2.1 Atlas selection

Instead of using all deformed label images we can make a selection of atlas scans and use only their associated deformed label images. The selection is based on the similarity of the patient image P and the deformed atlas images $A_i \circ T_i$. As in [64], we measure the similarity after registration by the normalised mutual information (NMI) [80]. Let us define the ratio r_i:

$$r_i = \frac{\mathrm{NMI}(P, A_i \circ T_i; \Omega)}{\max_j \mathrm{NMI}(P, A_j \circ T_j; \Omega)}. \tag{6.5}$$

An atlas A_i is selected if it satisfies $r_i \geq \varphi$, where $0 \leq \varphi \leq 1$ is a tunable parameter. A value of 0 means that all atlas scans are included in the selection. A value of 1 implies that only the atlas scan with the highest similarity measure is used. The settings $\varphi = 0$ and $\varphi = 1$ correspond to the "MUL" and "SIM" methods, respectively, investigated in [64]. In Sec. 6.4, we present results for a range of φ.

The set of atlas image indices selected in this stage is called $\mathcal{A}_P$. The subscript indicates that this set can be different for each patient image.

6.2.2.2 Majority voting (VOTE)

To combine the deformed segmentations of the selected atlas images into a single segmentation $L(\boldsymbol{x})$, majority voting is the most straightforward method. We consider a somewhat more general, weighted version, defined by the following two equations:

$$\ell_c(\boldsymbol{x}) = \frac{\sum_{i \in \mathcal{A}_P} w_i \cdot \delta\left[c, (L_i \circ T_i)(\boldsymbol{x})\right]}{\sum_{i \in \mathcal{A}_P} w_i}, \quad \forall c \in \mathcal{C}, \tag{6.6}$$

$$L(\boldsymbol{x}) = \arg\max_{c \in \mathcal{C}} \ell_c(\boldsymbol{x}), \tag{6.7}$$

where $\ell_c(\boldsymbol{x})$ denotes the probability of class c at $\boldsymbol{x}$, $\delta[\cdot]$ is the Kronecker delta function, and w_i are scalar weighting factors. Equation (6.7) selects the class with the highest probability as the final label. Setting $w_i = 1$ for all i yields the common majority voting procedure. By using $w_i = r_i$ more weight is assigned to atlas scans that match well to the patient image. Both approaches are tested in Sec. 6.4.

6.2.2.3 Simultaneous truth and performance level estimation (STAPLE)

The STAPLE algorithm [66, 94] treats label image fusion as a maximum-likelihood problem, which is solved using an expectation-maximisation (EM) procedure. Intuitively, the method is based on the following two observations: 1) if the patient segmentation L is known, the accuracy (reliability) of each deformed label image $L_i \circ T_i$ can be computed in terms of its specificity and sensitivity, and 2) if the specificity and sensitivity values of all deformed label images are known, a better estimate of L can be generated. In [66] it is demonstrated that the STAPLE algorithm gives better results than VOTE, when used for atlas-based segmentation of bee brains.

We run the STAPLE procedure using the *disputed* voxels only, i.e., the voxels where $(L_i \circ T_i)(\boldsymbol{x}) \neq (L_j \circ T_j)(\boldsymbol{x})$ for at least one combination of $i, j \in \mathcal{A}_P$. When the deformed label images are reasonably similar to each other, the disputed voxels lie on a narrow band around the prostate border. The STAPLE algorithm needs to be initialised with a probabilistic segmentation of each class. The probabilistic segmentation ℓ_c that results from VOTE, see Eq. (6.6), is a reasonable choice for this. The choice of w_i in Eq. (6.6) may influence the final STAPLE result, although the effect can be expected to be small, since the VOTE procedure serves here as an initialisation only. In Sec. 6.4 both $w_i = 1$ and $w_i = r_i$ are tested.

Note that, if all deformed label images indicate an over- or undersegmentation of the prostate, the final label image L will also be an over- or undersegmentation. This happens regardless of the label image fusion method (VOTE or STAPLE) and is not affected by the decision to use only the disputed voxels.

6.3 Experiments

Two data sets are available for the evaluation. The first set consists of 38 scans, originating from healthy volunteers. The second set consists of 50 clinical scans from prostate cancer patients.

For the experimental evaluation of the proposed segmentation method, an atlas needs to be defined. The composition of the atlas may have a large impact on the quality of the segmentations. The atlas should contain enough anatomical variation, such that for every target patient image a few atlas images are present that are reasonably similar to the patient image, allowing for successful registration. If the images are of very high quality, the diversity of the atlas may not be so important anymore, since the registration algorithm would match any pair of images successfully. We evaluate the influence of the atlas composition in our application by performing two types of experiments. In the first experiment the volunteer data set serves as an atlas and the clinical data set serves as a test set. The second experiment is a leave-one-out test, using only the patient data.

All experiments are performed both with MI and LMI as the similarity measure for registration. Also, the various atlas label fusion procedures described in Sec. 6.2.2 are tested. The results are evaluated by comparing the automatically generated segmentations with manual segmentations.

6.3.1 Data

6.3.1.1 Volunteer data

The volunteer data set consists of 38 MR scans, acquired with a Philips 3T scanner (Gyroscan NT Intera, Philips Medical Systems, Best, The Netherlands) using a flex-M coil and a balanced steady-state free precession (bSSFP) sequence with fat suppression. The scans originate from eight healthy volunteers (age 42-51 years, mean 47) and were made in the context of another study. Seven volunteers were scanned five times, one volunteer was scanned three times. The time between two

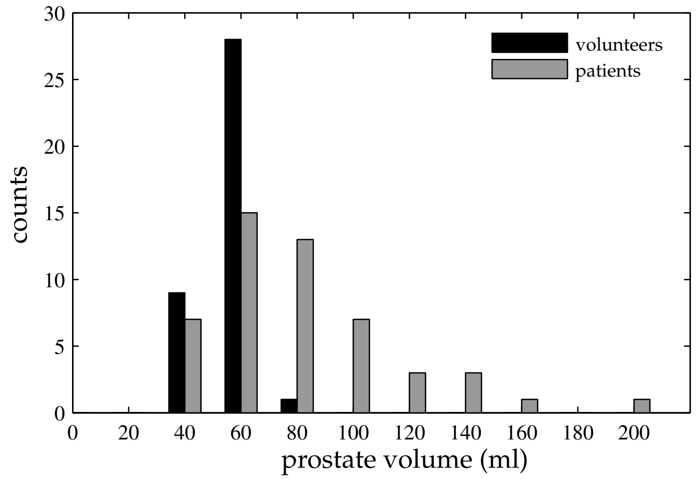


Figure 6.3: *Histogram of prostate volumes. The mean volume (± st.dev.) is 52 ± 6 ml for the volunteers and 82 ± 36 ml for the patients.*

scans was at least one day, and the volunteers were asked to try to vary the content of their rectum and bladder, to get as much variety between the scans as possible. The scans have a dimension of 512×512×90 voxels of size 0.49×0.49×1.0 mm. Manual segmentations are available for each scan. The segmentations were made by an experienced observer and approved by a radiation oncologist (observer A, see Sec. 6.3.1.2). Note that the seminal vesicles are considered part of the prostate. Besides the prostate, the bladder and the rectum were also labelled. The distribution of prostate volumes is visualised in Fig. 6.3 by the black bars of the histogram.

6.3.1.2 Patient data

The 50 clinical scans were acquired using the same protocol as the scans in the volunteer data set and originate from 50 prostate cancer patients (age 51-79 years, mean 69), which were scheduled for external beam radiation therapy. The patients did not have any loco-regional or distant metastases. For 35 patients the disease status was $T_{3,4}N_0M_0$. The rest was classified as $T_{1,2}N_0M_0$. In each scan the prostate was segmented by three observers. Observer A is a radiation oncologist and has the most experience (ten years) of the three observers. Observer B is a resident radiation oncologist and observer C is a medical physicist specialised in the field of prostate radiotherapy. We constructed an additional 'gold standard' L^G by combining the three segmentations L^A, L^B, and L^C using majority voting, with equal weights w_i. Note that only the prostate was delineated in the patient data. Bladder and rectum were not labelled. The distribution of prostate volumes as defined by L^G is shown in Fig. 6.3 by the grey bars. Clearly, a much larger range of prostate volumes is present in the patient data set than in the volunteer data set. It is well known that men in the age group of the patients often suffer from benign prostate hypertrophy (BPH). This can result in a substantial increase of the prostate volume.

6.3.2 Experiment description

6.3.2.1 Experiment I

In Experiment I the volunteer data set serves as an atlas and the clinical data set serves as a test set. For the label image fusion algorithms, VOTE and STAPLE, two choices of $\mathcal{C}$ (see Sec. 6.2.2) are considered: $\mathcal{C} = \{\text{background}, \text{prostate}\}$ and $\mathcal{C} = \{\text{background}, \text{prostate}, \text{rectum}, \text{bladder}\}$, where 'background' is defined as anything that does not belong to one of the other classes. The resulting label image fusion methods are referred to as VOTE2, VOTE4, STAPLE2, and STAPLE4, where the number indicates the number of classes in $\mathcal{C}$. Note that the rectum and bladder segmentations that come as a by-product from VOTE4 and STAPLE4 are not of our interest. We only assess the quality of the prostate delineation. As mentioned in Sec. 6.2.2, the presence in the atlas of additionally labelled tissue types besides prostate may improve the segmentation of the prostate in the target patient image.

During the registration of the atlas images to the patient images the similarity measure (MI or LMI) is evaluated on a region of interest Ω. A rectangular region of interest of $271 \times 333 \times 86$ voxels was manually selected for this purpose, roughly encompassing the prostate, bladder and rectum in all scans. For atlas selection, see Eq. (6.5), the same Ω is used.

6.3.2.2 Experiment II

The second experiment is a leave-one-out test, using only the patient data. For each patient the atlas set thus consists of the 49 remaining patients. The gold standard labels L^G are used as atlas label images. Only VOTE2 and STAPLE2 are considered in Experiment II, since no manual segmentations of the rectum and bladder are available for the patient data. For Ω the same definition is used as in Experiment I.

It may be expected that the results of Experiment II are better than those of Experiment I, since the atlas contains more anatomical variation, as shown in Fig. 6.3. We evaluate the relative impact of this difference in atlas composition, compared to other factors that influence the performance of the automatic segmentation method.

6.3.3 Evaluation measures

The results are evaluated by comparing the automatically generated prostate segmentations with the manual segmentations. A well-known measure of segmentation overlap is the Dice similarity coefficient (DSC) [18]:

$$\mathrm{DSC}(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|}, \tag{6.8}$$

where X and Y represent binary label images, and $|\cdot|$ denotes the number of voxels that equal 1. A higher DSC indicates a better correspondence. A value of 1 indicates perfect overlap, a value of 0 means no overlap at all.

The DSC does not provide insight in the spatial distribution of the segmentation errors. To visualise the segmentation accuracy we use a spherical coordinate mapping of the prostate boundary [22, 61]. The shortest Euclidean distance between the manual and automatic segmentation boundaries is computed for every point on the boundary of the manual segmentation. A cartographic 'Mollweide equal area' projection is used to display the result, as proposed in [22].

6.4 Results

The experiments were performed with MI and LMI, with VOTE2, VOTE4, STAPLE2, and STAPLE4, and with different thresholds for atlas selection (φ). The DSC values between the automatic segmentation L and the expert segmentations L^A, L^B, L^C, and L^G were computed for all 50 test scans. Figures 6.4 and 6.5 summarise the results of Experiment I and Experiment II, respectively. Each box-and-whisker in the lower parts of the figures visualises the distribution of $\text{DSC}(L, L^G)$ for a specific value of φ, when using LMI for the registration. The φ values can be found on the horizontal axis. The '-' symbol refers to $\varphi = 0$ combined with $w_i = 1$, i.e. no atlas selection and equal weights. In all other cases $w_i = r_i$ was used. Values of φ between 0 and 0.95 are not shown, since $\varphi = 0.95$ was already almost equivalent to $\varphi = 0$. An 'a' above the lower graph indicates significant improvement ($p < 0.05$, a value $p < 0.01$ never occurred) compared to $\varphi =$'-' with the same label fusion method. A 'b' indicates significant improvement ($p < 0.05$) compared to VOTE2 with the same value of φ. Statistical significance was evaluated using a paired two-sided Wilcoxon test. The upper parts of Figs. 6.4 and 6.5 show the effect of LMI compared to MI. Each group of four bars displays the medians of the differences $\text{DSC}(L^{\text{LMI}}, L^E) - \text{DSC}(L^{\text{MI}}, L^E)$, for $L^E \in \{L^A, L^B, L^C, L^G\}$. Grey bars indicate that the difference is significant according to a paired two-sided Wilcoxon statistical test ($p < 0.01$).

The upper parts of the figures clearly show that LMI outperformed MI in this application. The median DSC difference was positive (favouring LMI) for all settings of φ, with all tested label image fusion methods, both in Experiment I and Experiment II. Also, the choice of ground truth (L^A, L^B, L^C, or L^G) did not change the conclusion. Almost all differences were significant with $p < 0.01$. Only in combination with STAPLE4 the difference was not always significant. However, the lower graph shows that STAPLE4 produced the worst results of all label image fusion methods in Experiment I. The advantage of LMI comes at no additional computational costs, thanks to the stochastic optimisation method, as explained in Sec. 6.2.1. The measure computation time was around 15 minutes per registration on a single processor Pentium 2.8GHz personal computer. For the implementation of LMI in [78] a computation time of 1-2 hours per registration is reported. In [24] the authors report a computation time of 30 minutes on a cluster of 24 processors.

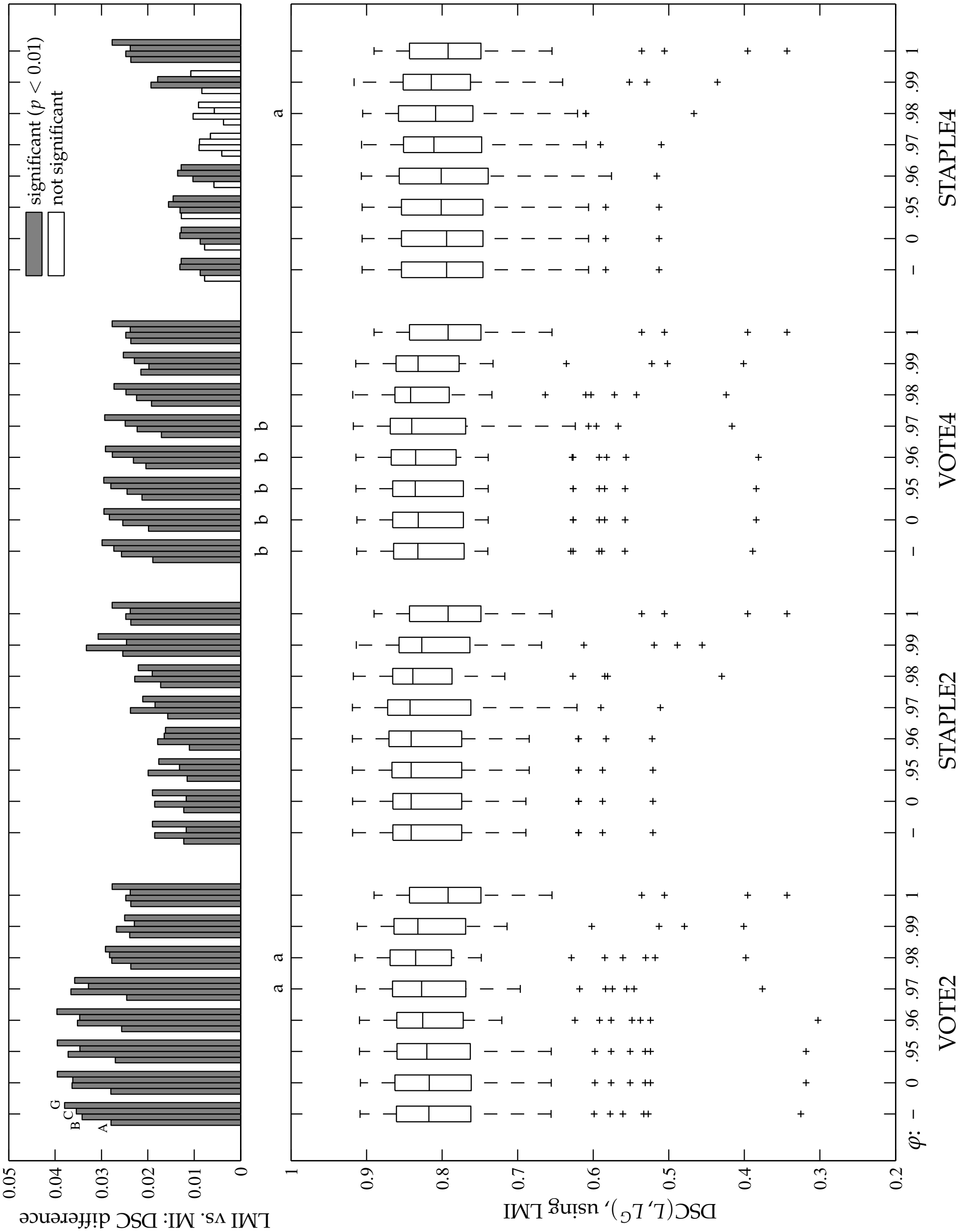

LMI vs. MI: DSC difference
0.05
0.04
0.03
0.02
0.01
0
significant ($p < 0.01$)
not significant
A
B
C
G
a a
b b b b
a
DSC(L, L^G), using LMI
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
φ: − 0 .95 .96 .97 .98 .99 1
VOTE2
STAPLE2
VOTE4
STAPLE4

Figure 6.4: *(previous page) Results for Experiment I. The lower part of the figure shows the effect of the four label image fusion methods, each for a range of values of φ, using LMI as the similarity measure. The upper part of the graph visualises the difference between using LMI and MI. An 'a' above the lower graph indicates significant ($p < 0.05$) improvement compared to φ = '-' with the same label fusion method. A 'b' indicates significant ($p < 0.05$) improvement compared to VOTE2 with the same value of φ.*

The lower graphs in Figs. 6.4 and 6.5 show that the differences between the different label image fusion methods were mostly rather small, but statistically significant in some cases. Selecting only the most similar atlas ($\varphi = 1$) gave the worst results, which confirms the results found in [64]. The optimal value of φ was around 0.98, for both the VOTE and STAPLE methods. With this value, on average 22 out of 38 atlas images were selected in Experiment I and 23 out of 49 in Experiment II. In contrast to the results reported in [66], the STAPLE algorithm did not clearly improve upon the VOTE method in our application. In Experiment I, STAPLE2 yielded somewhat better results than VOTE2 for $\varphi < 0.96$, but the difference was not statistically significant. For higher values of φ, STAPLE2 and VOTE2 performed equally. STAPLE4 gave worse results than VOTE4 for all φ. In Experiment II, for all values of φ, STAPLE2 performed slightly worse than VOTE2. The additionally labelled structures in the atlas set (rectum and bladder) taken into account by VOTE4 and STAPLE4 did not lead to consistently better results either. VOTE4 improved slightly upon VOTE2, but the difference was significant only for $\varphi < 0.98$. STAPLE4 performed worse than STAPLE2 for all φ. For further evaluation we use VOTE2 with $\varphi = 0.98$ and LMI as the similarity measure.

Figure 6.6 compares the automatic segmentation results with the interobserver variability. Each box-and-whisker visualises the distribution of DSC values over all 50 patients. For Experiment I and II the distributions of $\mathrm{DSC}(L, L^E)$ for $L^E \in \{L^A, L^B, L^C, L^G\}$ are shown. In Experiment I, the median DSC varied between 0.80 (with L^C) and 0.84 (with L^G). In Experiment II, the median varied between 0.85 (with L^A, L^B, and L^C) and 0.88 (with L^G). When comparing the results of Experiment I and Experiment II, the large impact of the atlas composition becomes clear. The median differences between the corresponding DSC values of Experiment I and II were all significant with $p < 0.001$, according to a paired two-sided Wilcoxon statistical test. The interobserver overlap values $\mathrm{DSC}(L^A, L^B)$, $\mathrm{DSC}(L^A, L^C)$, and $\mathrm{DSC}(L^B, L^C)$ had median values of around 0.87. The results of Experiment II thus approached the level of the interobserver variability, although the human observers remained superior. The overlap of the expert segmentations with L^G was highest, which is not surprising, since L^G was constructed from L^A, L^B, and L^C by majority voting. Among the three experts, observer A had the highest median DSC with L^G. Based on this and on the fact that observer A was the most experienced one, we choose to use L^A as a ground truth in the following analysis of the spatial distribution of segmentation errors.

Figure 6.8 shows the spatial distribution of the segmentation errors. A "Moll-

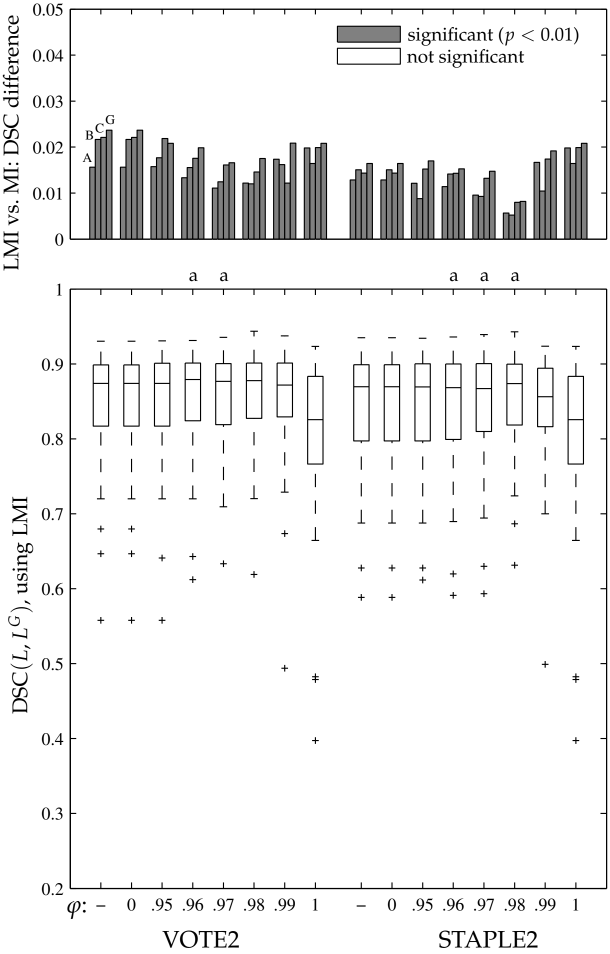


Figure 6.5: *Results for Experiment II. The lower part of the figure shows the effect of the two label image fusion methods, each for a range of values of φ, using LMI as the similarity measure. The upper part of the graph visualises the difference between using LMI and MI. An 'a' above the lower graph indicates significant ($p < 0.05$) improvement compared to φ = '-' with the same label fusion method. A 'b' would indicate significant ($p < 0.05$) improvement compared to VOTE2 with the same value of φ, but this situation never occurred.*

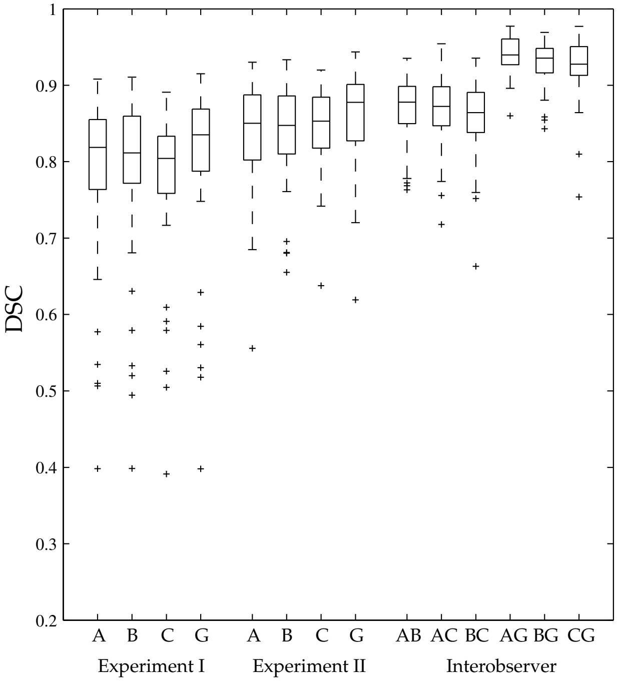


Figure 6.6: *The results of Experiment I and II compared to the interobserver variability. The automatic segmentation results shown were generated using LMI, VOTE2, and $\varphi = 0.98$.*

weide" map of the prostate surface is given in Fig. 6.7. For each test scan, the shortest Euclidean distance between the boundaries of L and L^A was computed at every point on the boundary of L^A. Subsequently, the computed distances were projected on the Mollweide map. The results of the 50 test scans were summarised by computing the first quartile, median, and third quartile of the distance at every location on the Mollweide map. Figures 6.8(a)-(c) and 6.8(d)-(f) show the results for Experiment I and Experiment II, respectively. In order to assess the interobserver variation, the distances between L^B and L^A and between L^C and L^A were also calculated. These results are shown in Figs. 6.9. Note that different colour scales are used for the first quartile, median, and third quartile plots.

From the figures it is evident that in Experiment I the largest errors occurred at the border between the prostate and the bladder. In Experiment II the errors at the prostate-bladder boundary were much smaller, and were even comparable to the interobserver distance between L^B and L^A. The most serious segmentation errors in Experiment II were made in the tips of the seminal vesicles, which was confirmed by visual inspection of the segmentations. Both for the automatic segmentations and for the experts, the errors at the apex were relatively high. Somewhat larger er-

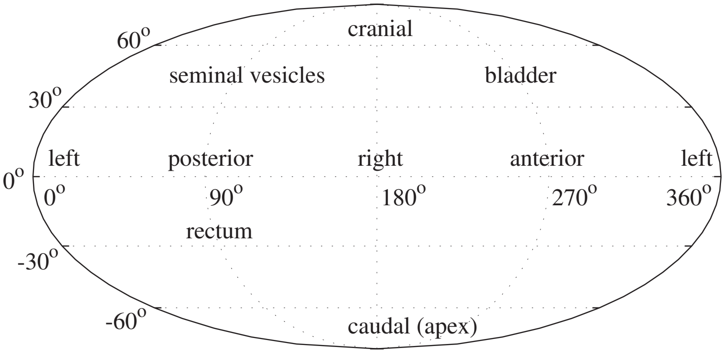


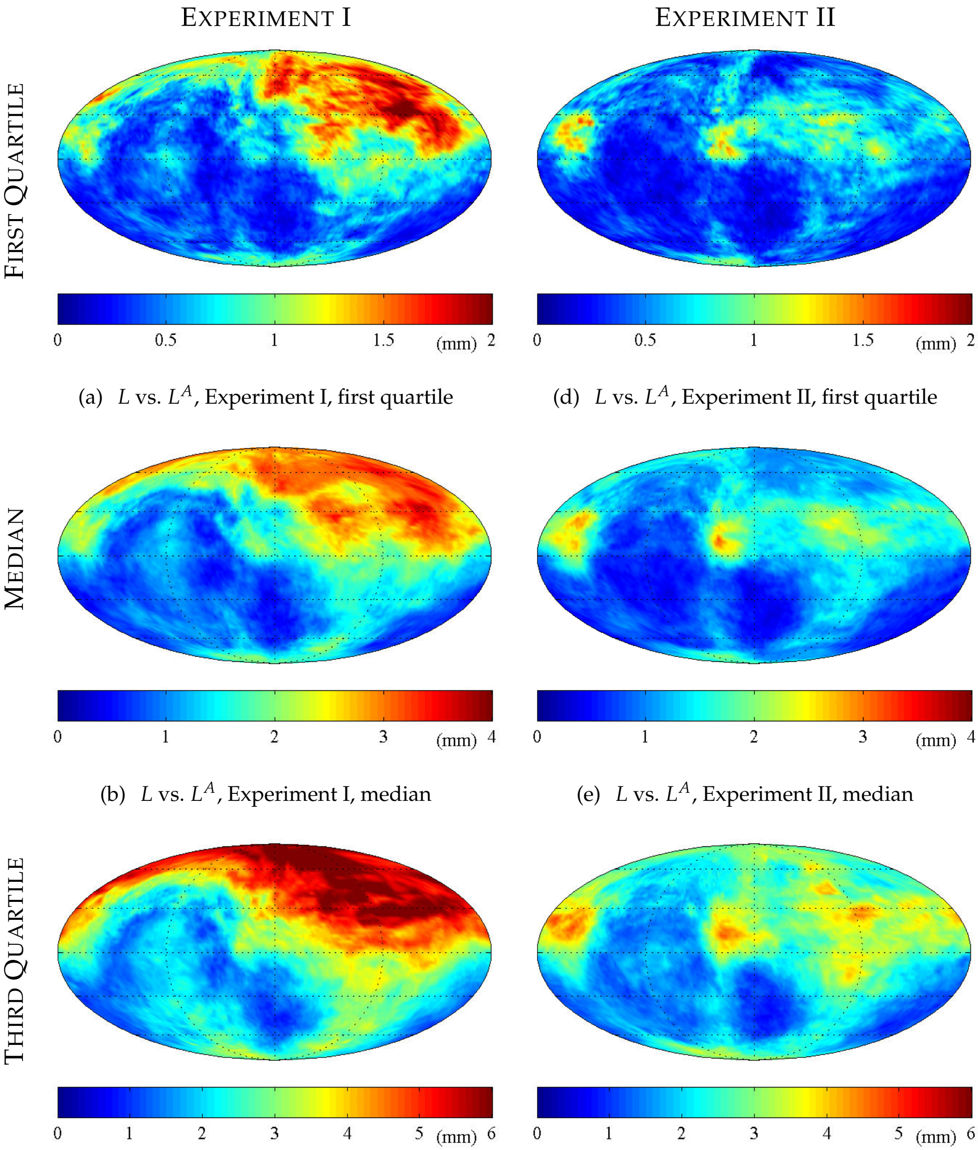
Figure 6.7: *Mollweide projection of the prostate boundary.*

rors were also observed at the anterior side of the prostate. At the prostate-rectum interface observer B and C were very close to observer A; Figures 6.9(c) and 6.9(f) show that in 75% of the cases the deviation remained below 1 mm. In Experiment II, the automatic segmentation errors at the prostate-rectum interface remained below 1 mm in 50% of the cases and below 1.5 mm in 75% of the cases, as shown in Figs. 6.8(e) and 6.8(f), respectively.

6.5 Discussion

The accuracy of the automatic segmentation method is on a large part of the prostate surface close to the level of interobserver variability, for most test images. Most serious errors occurred around the tips of the seminal vesicles and at the anterior side of the prostate. The automatic method showed especially good performance at the prostate-rectum interface, although the human observers remained superior in most cases. Whereas a segmentation error of a few millimetres is clinically acceptable at boundaries with muscular tissue, the interfaces with rectum and bladder need to be delineated with an accuracy equal to the level of interobserver variability. Further improvement of the method is, thus, necessary.

Visual inspection revealed that the large segmentation errors at the prostate-bladder boundary in Experiment I (see Fig. 6.8) mainly occurred when the patient's prostate was very large. The volunteer data set, which is used as an atlas in Experiment I, does not contain any examples of large prostates, as shown in Fig. 6.3. Matching the atlas images to the patient image is thus likely to fail. A large number of the outliers observed in Fig. 6.6 for Experiment I can be attributed to this. The large differences between the results of Experiment I and II emphasise the importance of a proper atlas composition. Therefore, we expect that the automatic segmentation results can be further improved by explicit optimisation of the atlas composition in an initial training procedure.



(a) L vs. L^A, Experiment I, first quartile

(d) L vs. L^A, Experiment II, first quartile

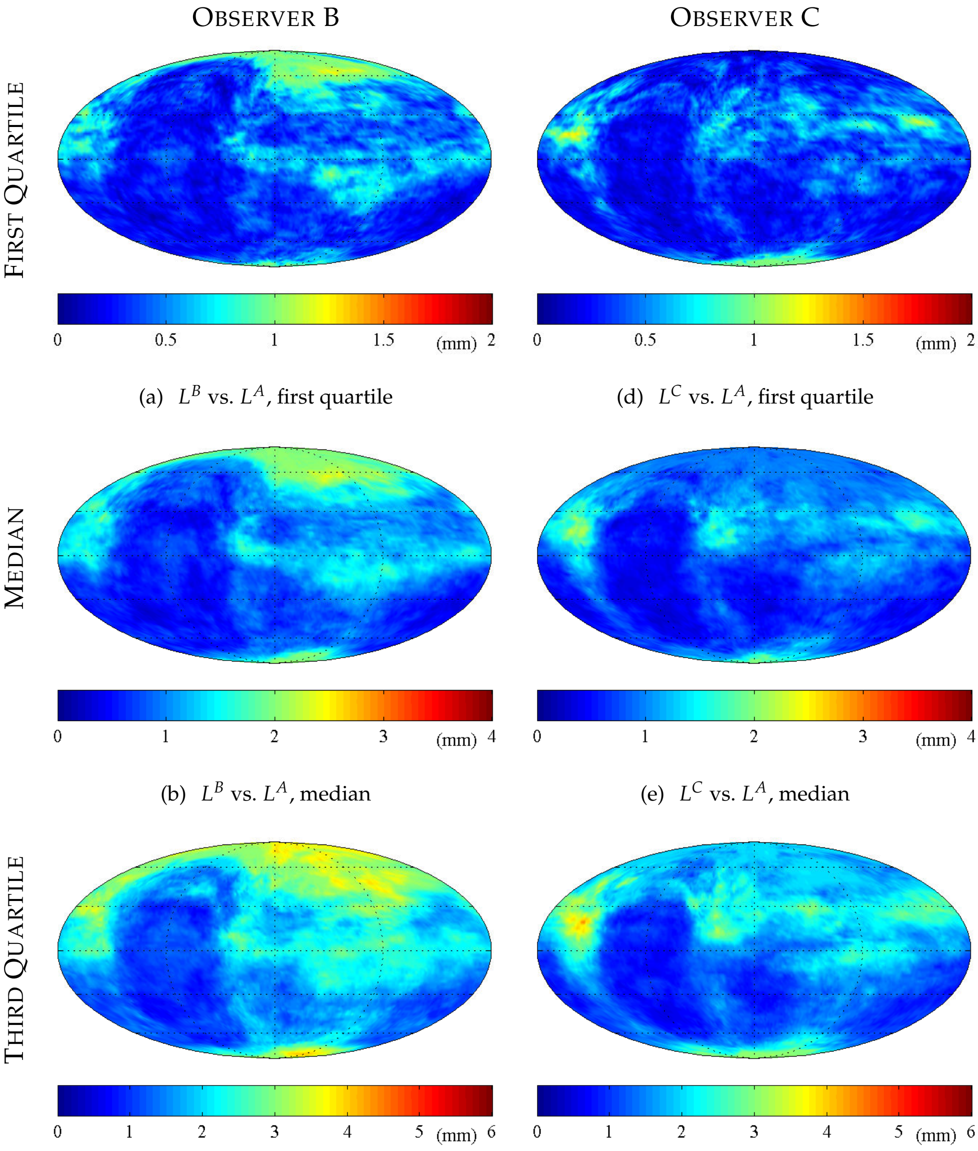
(b) L vs. L^A, Experiment I, median

(e) L vs. L^A, Experiment II, median

(c) L vs. L^A, Experiment I, third quartile

(f) L vs. L^A, Experiment II, third quartile

Figure 6.8: *The spatial distribution of automatic segmentation errors. Figure 6.7 shows a map of the prostate surface that aids in interpretation. Note that the graphs have different colour scales. The segmentation results shown were generated using LMI, VOTE2, and* $\varphi = 0.98$.



(a) L^B vs. L^A, first quartile

(d) L^C vs. L^A, first quartile

(b) L^B vs. L^A, median

(e) L^C vs. L^A, median

(c) L^B vs. L^A, third quartile

(f) L^C vs. L^A, third quartile

Figure 6.9: *The spatial distribution of the interobserver variation (with L^A as reference). Figure 6.7 shows a map of the prostate surface that aids in interpretation. Note that the graphs have different colour scales.*

In previous work [33], we have investigated to put more weight on the prostate region during registration, as a possible way to improve the results, by defining a narrow region of interest around the prostate segmentation in the atlas scan, and use only that part for registration. Experiments on the volunteer data set (used also in the current manuscript) showed some improvement over standard MI registration, but the use of LMI showed superior results on the same data set [34].

The MR images used to evaluate our algorithm were acquired using a bSSFP sequence with fat suppression. Using this sequence, a high resolution ($0.49 \times 0.49 \times 1.0$ mm) is obtained in a scan time of about two minutes on a 3T MR scanner. The protocol was optimised for maximum contrast, to facilitate manual prostate contouring. It remains to be investigated whether the protocol is optimal for *automatic* prostate segmentation. The bSSFP sequence is sensitive to susceptibility artefacts, for instance, due to the presence of air in the rectum. These artefacts disturbed the nonrigid registration procedure in some cases.

The clinical test images from the patient data set are planning scans of patients scheduled for *external beam* radiation therapy. In clinical practice, the calculations of the dose distributions are actually based on these scans. In case of *brachytherapy*, the dose plan may be adapted based on intra-operative scans. We expect that the deformations induced by the insertion of seeds do not form a problem, given the already large anatomical variability that the method is able to cope with, and the varying states of rectum and bladder filling. Susceptibility artefacts in the images due to the presence of the seeds will become an additional challenge though.

To the best of our knowledge, no other automatic segmentation results for the prostate including the seminal vesicles on 3D MR scans are available in the literature. In [54] a semi-automatic segmentation method is presented for the prostate without seminal vesicles. The method is evaluated on 3D MR scans of 24 patients. The results are given in terms of the "volume overlap" between manual and automatic segmentations. The volume overlap is also known as the Tannimoto coefficient (TC) and is related to the DSC by $\text{DSC} = 2\text{TC}/(\text{TC}+1)$ [12]. A mean TC of 0.78 is reported with standard deviation ± 0.05, which corresponds to a DSC of 0.88 ± 0.04. This is somewhat better than our results in Fig. 6.6 for Experiment II. However, the presence of the seminal vesicles increases the surface-to-volume ratio of the segmented structure, which increases the sensitivity of the DSC measure [64]. In [98], a pseudo-3D active shape modelling approach is used to segment the prostate without seminal vesicles. The method is validated on 26 3D MR scans, on a slice-by-slice basis, using the root mean square distance (RMSD) between the manual and automatic segmentation. A mean RMSD of 5.5 mm with a standard deviation of ± 2.9 mm is reported. We may compare this result to Fig. 6.8(f), which shows that, with our method, the segmentation error remained at every location below 5 mm in 75% of the test cases. While this result seems to be in favour of our method, the RMSD values reported in [98] might be lower if they would not have been computed on a slice-by-slice basis, but in 3D. Also, it should be noted that both methods mentioned above [54, 98] were validated on MR scans, acquired using a 1.5T machine, with highly anisotropic voxels.

6.6 Conclusion

An automatic prostate segmentation method for pelvic MR images has been proposed. The method is based on matching of manually segmented atlas images. To account for the large variability in shape, multiple atlas images are combined. A computationally efficient localised mutual information similarity measure is used in the matching stage. Evaluation was performed on a set of 50 clinical scans, which were manually segmented by three experts.

The choice of similarity measure and the composition of the atlas were demonstrated to be important determinants of segmentation quality. Using localised mutual information instead of standard mutual information yielded a significant ($p < 0.01$) improvement of around 0.02, in terms of the median Dice similarity coefficient (DSC). Using an atlas composed of patient data instead of volunteer data resulted in a median DSC increase of 0.04 (significant with $p < 0.01$), accompanied by a great reduction of the number of outliers. The label image fusion procedure had only a modest influence on the results. A majority voting method with an atlas selection level of $\varphi = 0.98$ gave good results.

With the best settings, a median DSC of about 0.85 was achieved for the prostate, which is close to the interobserver variability of 0.87. The segmentation quality was especially good at the prostate-rectum interface, where the segmentation error remained below 1 mm in 50% of the cases and below 1.5 mm in 75% of the cases.

Chapter 7

Summary and Discussion

Image registration is an important task in medical image processing. It refers to the process of aligning data sets, possibly from different modalities (e.g., magnetic resonance (MR) and computed tomography (CT)), different time points (e.g., follow-up scans), and/or different subjects (in case of population studies). The registration problem can mathematically be formulated as an optimisation problem. Methods for solving this optimisation problem are studied in this thesis. Below, a summary of the thesis is given:

Chapter 2 A large number of methods for image registration are described in the literature. Unfortunately, there is not one method that works for all applications. This chapter presents `elastix`: a publicly available computer program for intensity-based medical image registration. The software consists of a collection of algorithms that are commonly used to solve medical image registration problems. The modular design of `elastix` allows the user to quickly configure, test, and compare different registration methods for a specific application. The command-line interface enables automated processing of large numbers of data sets, by means of scripting. The usage of `elastix` for comparing different registration methods is illustrated with three example experiments, in which individual components of the registration method are varied. The `elastix` software is used for all experiments described in this thesis.

Chapter 3 A popular technique for nonrigid registration of medical images is based on the maximisation of their mutual information, in combination with a deformation field parameterised by cubic B-splines. The coordinate mapping that relates the two images is found using an iterative optimisation procedure. This chapter compares the performance of eight optimisation methods: gradient descent (with two different step size selection algorithms), quasi-Newton, nonlinear conjugate gradient, Kiefer-Wolfowitz, simultaneous perturbation, Robbins-Monro (RM), and evolution strategy. Special attention is paid to computation time reduction by using fewer voxels to calculate the cost function and its derivatives. The optimisation methods are tested on manually deformed CT images of the heart, on follow-up CT chest scans, and on MR scans of the prostate acquired using a BFFE, T1, and T2 protocol. Registration accuracy is assessed by computing the overlap of segmented edges. Precision and convergence properties are studied by comparing deformation fields. The results show that the RM method is the best choice in most applications. With this approach, the computation time per iteration can be lowered approximately 500 times without affecting the rate of convergence by using a small subset of the image, randomly selected in every iteration, to compute the derivative of the mutual information. From the other methods the quasi-Newton and the nonlinear conjugate gradient method achieve a slightly higher precision, at the price of larger computation times.

Chapter 4 This chapter presents a stochastic optimisation method for intensity-based monomodal image registration. The method is based on the RM stochastic gradient descent method and adds a preconditioning matrix. The derivation of the preconditioner is based on the observation that, after registration, the deformed moving image should approximately equal the fixed image. This prior knowledge allows us to approximate the Hessian at the minimum of the registration cost function, without knowing the coordinate transformation that corresponds to this minimum. The method is validated using 3D functional MRI time-series and 3D CT chest follow-up scans. The experimental results show that the preconditioned stochastic gradient descent method (PSGD) accelerates convergence and simplifies parameter selection, in comparison with Robbins-Monro.

Chapter 5 This chapter present a stochastic gradient descent optimisation method for image registration with adaptive step size prediction. The method is based on theoretical work by Plakhov and Cruz. Our main methodological contribution is the derivation of an image-driven mechanism to select proper values for the most important free parameters of the method. The selection mechanism employs general characteristics of the cost functions that commonly occur in intensity-based image registration. Also, the theoretical convergence conditions of the optimisation method are taken into account. The proposed adaptive stochastic gradient descent (ASGD) method is compared to a standard, non-adaptive RM algorithm. Both ASGD and RM employ a stochastic subsampling technique to accelerate the optimisation process. Registration experiments were performed on 3D CT and MR data of the head, lungs, and prostate, using various similarity measures and transformation models. The results indicate that ASGD is robust to these variations in the registration framework and is less sensitive to the settings of the user-defined parameters than RM. The main disadvantage of RM is the need for a predetermined step size function. The ASGD method provides a solution for that issue.

Chapter 6 An automatic method for delineating the prostate (including the seminal vesicles) in 3D MR scans is presented in this chapter. The method is based on nonrigid registration of a set of prelabelled atlas images. Each atlas image is nonrigidly registered with the target patient image. Subsequently, the deformed atlas label images are fused to yield a single segmentation of the patient image. The proposed method is evaluated on 50 clinical scans, which were manually segmented by three experts. The Dice similarity coefficient (DSC) is used to quantify the overlap between the automatic and manual segmentations. We investigate the impact of several factors on the performance of the segmentation method. For the registration, two similarity measures are compared: mutual information and a localised version of mutual information. The latter turns out to be superior (median $\Delta\text{DSC} \approx 0.02$, $p < 0.01$ with a paired two-sided Wilcoxon test) and comes at no added computational cost, thanks to the use of a stochastic optimisation scheme based on the principles behind the RM method. For the atlas fusion step we consider a majority voting rule and the "simultaneous truth and performance level estimation"

(STAPLE) algorithm, both with and without a preceding atlas selection stage. The differences between the various fusion methods appear to be small and mostly not statistically significant ($p > 0.05$). To assess the influence of the atlas composition, two atlas sets are compared. The first set consists of 38 scans of healthy volunteers. The second set is constructed by a leave-one-out approach using the 50 clinical scans that are used for evaluation. The second atlas set gives substantially better performance ($\Delta\text{DSC} = 0.04$, $p < 0.01$), stressing the importance of a careful atlas definition. With the best settings, a median DSC of around 0.85 is achieved, which is close to the median interobserver DSC of 0.87. The segmentation quality is especially good at the prostate-rectum interface, where the segmentation error remains below 1 mm in 50% of the cases and below 1.5 mm in 75% of the cases.

In the introduction of this thesis (Chapter 1) the following research questions were formulated:

- Which optimisation method to use?
- How can we estimate reasonable values for the user-defined parameters (if any) of the optimisation method?

The experimental results described in this thesis indicate that the RM stochastic gradient descent method, or one of its enhanced versions PSGD and ASGD, is a good choice in many applications. The stochastic subsampling technique employed by these methods substantially reduces the computation time per iteration, while convergence properties are retained. The basic RM method has the drawback that it requires the user to set some important parameters, related to the step size in each iteration. The methods PSGD and ASGD were introduced to solve this issue. The PSGD method is aimed at monomodal registration problems. The ASGD method is more generally applicable, but makes some simplifying assumptions that may not always be satisfied.

Of course the experiments are not exhaustive. It is impossible to claim that the conclusions hold in all situations. However, we demonstrated good results in a large variety of registration problems. Tests were performed using rigid, affine, and nonrigid transform models, with various cost functions (mean squared difference, normalised correlation, mutual information, normalised mutual information, localised mutual information), different image modalities (CT, MR), and several anatomical structures (heart, lungs, prostate, brain). The `elastix` software described in Chapter 2 enables the reader to extend the experiments to his/her own application.

This thesis is aimed at *medical* image registration, as suggested by the title. The fact that medical images are involved raises the question: "What is the clinical impact of this study?". The results presented in this thesis clearly demonstrate that the choice of optimisation method can have a large impact on computation time, robustness, and accuracy, which are all relevant aspects in clinical practice. It is shown how accurate nonrigid registration of large 3D image volumes can be achieved within minutes. The automatic prostate contouring method presented in Chapter 6 is an example of an application where fast registration is a necessary prerequisite. Around 50 registrations are used for the segmentation of a single prostate. This would not be feasible with a registration algorithm that takes hours to complete.

By making the source code of the algorithms publicly available as a part of `elastix`, it becomes possible for researchers in the field of medical imaging to directly incorporate the methods in their own applications. We hope that the algorithms presented in this thesis thus will find their way to clinical practice.

Acknowledgements

This work benefited from the use of the Insight Segmentation and Registration Toolkit (ITK), an open source software package developed as an initiative of the U.S. National Library of Medicine and available at http://www.itk.org. The brain images used in Chapters 2 and 5 originated from the "Retrospective Image Registration Evaluation" project, National Institutes of Health, Project Number 8R01EB002124-03, Principal Investigator, J. Michael Fitzpatrick, Vanderbilt University, Nashville, TN. The Mollweide projections in Chapter 6 were generated using the "M_Map" package, developed by Rich Pawlowicz, which is available at http://www.eos.ubc.ca/~rich. We thank the authors of the "Matrix Cookbook" [57], which has been a valuable resource.

References

[1] P. Anbeek, K.L. Vincken, M.J.P. van Osch, R.H.C. Bisschops, and J. van der Grond. Automatic segmentation of different-sized white matter lesions by voxel probability estimation. *Medical Image Analysis*, 8(3):205–215, 2004.

[2] D.V. Arnold and H.-G. Beyer. Evolution strategies with cumulative step length adaptation on the noisy parabolic ridge. *Natural Computing*, in press.

[3] V. Arsigny, X. Pennec, and N. Ayache. Polyrigid and polyaffine transformations: a novel geometrical tool to deal with non-rigid deformations - application to the registration of histological slices. *Medical Image Analysis*, 9(6):507–523, 2005.

[4] A. Auger. Convergence results for the $(1, \lambda)$-SA-ES using the theory of ϕ-irreducible markov chains. *Theoretical Computer Science*, 334:35–69, 2005.

[5] R. Bajcsy and S. Kovačič. Multiresolution elastic matching. *Computer Vision, Graphics and Image Processing*, 46(1):1–21, 1989.

[6] D.P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 2nd edition, 1999.

[7] H.-G. Beyer and S. Meyer-Nieberg. Self-adaptation of evolution strategies under noisy fitness evaluations. *Genetic Programming and Evolvable Machines*, 7:295–328, 2006.

[8] H.-G. Beyer and H.-P. Schwefel. Evolution strategies - a comprehensive introduction. *Natural Computing*, 1(12):3–52, 2002.

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

[10] G.E. Christensen and H.J. Johnson. Consistent image registration. *IEEE Transactions on Medical Imaging*, 20(7):568–582, 2001.

[11] A.A. Cole-Rhodes, K.L. Johnson, J. LeMoigne, and I. Zavorin. Multiresolution registration of remote sensing imagery by optimization of mutual information using a stochastic gradient. *IEEE Transactions on Image Processing*, 12(12):1495–1511, 2003.

[12] W.R. Crum, O. Camara, and D.L.G. Hill. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Transactions on Medical Imaging*, 25(11):1451–1461, 2006.

[13] P. Cruz. Almost sure convergence and asymptotical normality of a generalization of Kesten's stochastic approximation algorithm for multidimensional case. Technical report, Cadernos de Matemática, Série de Investigação, Collection of University of Aveiro, Department of Mathematics, 2005. `http://193.136.81.248/dspace/handle/2052/74`.

[14] P. Cruz. *Aproximação Estocástica com Valor do Passo Adaptativo*. PhD thesis, University of Aveiro, Department of Mathematics, 2005. `http://193.136.81.248/dspace/handle/2052/103`.

[15] Y.-H. Dai. An efficient hybrid conjugate gradient method for unconstrained optimization. *Annals of Operations Research*, 103:33–47, 2001.

[16] Y.-H. Dai. A family of hybrid conjugate gradient methods for unconstrained optimization. *Mathematics of Computation*, 72(243):1317–1328, 2003.

[17] J.E. Dennis Jr. and J.J. Moré. Quasi-Newton methods, motivation and theory. *SIAM Review*, 19(1):46–89, 1977.

[18] L.R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[19] V. Fabian. On asymptotic normality in stochastic approximation. *Annals of Mathematical Statistics*, 39(4):1327–1332, 1968.

[20] B. Fei, J.L. Duerk, D.B. Sodee, and D.L. Wilson. Semiautomatic nonrigid registration for the prostate and pelvic MR volumes. *Academic Radiology*, 12(7):815–824, 2005.

[21] B. Fischer and J. Modersitzki. A unified approach to fast image registration and a new curvature based registration technique. *Linear Algebra Applications*, 380:107–124, 2004.

[22] M. Foskey, B. Davis, L. Goyal, S. Chang, E. Chaney, N. Strehl, S. Tomei, J. Rosenman, and S. Joshi. Large deformation three-dimensional image registration in image-guided radiation therapy. *Physics in Medicine and Biology*, 50:5869–5892, 2005.

[23] N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

[24] G. Hermosillo. *Variational Methods for Multimodal Image Matching*. PhD thesis, Université de Nice - Sophia Antipolis, 2002. `http://rangiroa.essi.fr/riveill/rapports/2002/02-these-hermosillo.pdf`.

[25] M.R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49:409–436, 1952.

[26] D.L.G. Hill, P.G. Batchelor, M. Holden, and D.J. Hawkes. Medical image registration. *Physics in Medicine and Biology*, 46(3):R1–R45, 2001.

[27] S. Hu, E.A. Hoffman, and J.M. Reinhardt. Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images. *IEEE Transactions on Medical Imaging*, 20(6):490–498, 2001.

[28] L. Ibáñez, W. Schroeder, L. Ng, and J. Cates. *The ITK Software Guide*. Kitware, Inc. ISBN 1-930934-15-7, second edition, 2005.

[29] L.O. Jay. A note on Q-order of convergence. *BIT Numerical Mathematics*, 41(2):422–429, 2001.

[30] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, 23(3):462–466, 1952.

[31] S. Klein, M. Staring, and J.P.W. Pluim. Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines. *IEEE Transactions on Image Processing*, 16(12):2879–2890, 2007.

[32] S. Klein, U.A. van der Heide, I.M. Lips, M. van Vulpen, M. Staring, and J.P.W. Pluim. Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information. *Medical Physics*, 35(4):1407–1417, 2008.

[33] S. Klein, U.A. van der Heide, B.W. Raaymakers, A.N.T.J. Kotte, M. Staring, and J.P.W. Pluim. Segmentation of the prostate in MR images by atlas matching. In J.A. Fessler and T.S. Denney Jr., editors, *4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1300–1303, 2007.

[34] S. Klein, U.A. van der Heide, M. Staring, A.N.T.J. Kotte, B.W. Raaymakers, and J.P.W. Pluim. Segmentation of the prostate in MR images by atlas matching using localised mutual information. In D.A. Jaffray, M. Sharpe, J. van Dyk, and J.P. Bissonnette, editors, *XVth International Conference on the use of Computers in Radiation Therapy*, volume 2, pages 585–589, 2007.

[35] H.J. Kushner and D.S. Clark. *Stochastic approximation methods for constrained and unconstrained systems*. Springer-Verlag, New York, 1978.

[36] H.J. Kushner and G.G. Yin. *Stochastic approximation and recursive algorithms and applications*. Springer-Verlag, New York, second edition, 2003.

[37] J. Kybic and M. Unser. Fast parametric elastic image registration. *IEEE Transactions on Image Processing*, 12(11):1427–1442, 2003.

[38] M. Lefébure and L.D. Cohen. Image registration, optical flow and local rigidity. *Journal of Mathematical Imaging and Vision*, 14(2):131–147, 2001.

[39] H. Lester and S.R. Arridge. A survey of hierarchical non-linear medical image registration. *Pattern Recognition*, 32(1):129–149, 1999.

[40] B. Likar and F. Pernuš. A hierarchical approach to elastic registration based on mutual information. *Image and Vision Computing*, 19(1-2):33–44, 2001.

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

[42] F. Maes, D. Vandermeulen, and P. Suetens. Comparative evaluation of multiresolution optimization strategies for multimodality image registration by maximization of mutual information. *Medical Image Analysis*, 3(4):373–386, 1999.

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

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

[45] K. Mathiak and S. Posse. Evaluation of motion and realignment for functional magnetic resonance imaging in real time. *Magnetic Resonance in Medicine*, 45:167–171, 2001.

[46] D. Mattes, D.R. Haynor, H. Vesselle, T.K. Lewellen, and W. Eubank. PET-CT image registration in the chest using free-form deformations. *IEEE Transactions on Medical Imaging*, 22(1):120–128, 2003.

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

[48] J. Modersitzki. *Numerical Methods for Image Registration*. Oxford University Press, 2004.

[49] J.J. Moré and D.J. Thuente. Line search algorithms with guaranteed sufficient decrease. *ACM Transactions on Mathematical Software*, 20(3):286–307, 1994.

[50] H. Mukai. Readily implementable conjugate gradient methods. *Mathematical Programming*, 17:298–319, 1979.

[51] K. Murphy, B. van Ginneken, J.P.W. Pluim, S. Klein, and M. Staring. Semi-automatic reference standard construction for quantitative evaluation of lung CT registration. In *Proceedings of MICCAI 2008*, in press.

[52] J. Nocedal. Updating quasi-Newton matrices with limited storage. *Mathematics of Computation*, 35(151):773–782, 1980.

[53] J. Nocedal and S.J. Wright. *Numerical optimization*. Springer-Verlag, New York, 1999.

[54] D. Pasquier, T. Lacornerie, M. Vermandel, J. Rousseau, E. Lartigau, and N. Bertrouni. Automatic segmentation of pelvic structures from magnetic resonance images for prostate cancer radiotherapy. *International Journal of Radiation Oncology, Biology, Physics*, 68(2):592–600, 2007.

[55] X. Pennec, P. Cachier, and N. Ayache. Tracking brain deformations in time sequences of 3D US images. *Pattern Recognition Letters*, 24(4-5):801–813, 2003.

[56] G.P. Penney, J. Weese, J.A. Little, P. Desmedt, D.L.G. Hill, and D.J. Hawkes. A comparison of similarity measures for use in 2D-3D medical image registration. *IEEE Transactions on Medical Imaging*, 17(4):586–595, 1998.

[57] K.B. Petersen and M.S. Pedersen. *The Matrix Cookbook*. `http://matrixcookbook.com`, September 2007.

[58] A. Plakhov and P. Cruz. A stochastic approximation algorithm with step size adaptation. *Journal of Mathematical Sciences*, 120(1):964–973, 2004.

[59] J.P.W. Pluim, J.B.A. Maintz, and M.A. Viergever. Interpolation artefacts in mutual information-based image registration. *Computer Vision and Image Understanding*, 77(2):211–232, 2000.

[60] F.A. Potra and Y. Shi. Efficient line search algorithm for unconstrained optimization. *Journal of Optimization Theory and Applications*, 85(3):677–704, 1995.

[61] C. Rasch, I. Barillot, P. Remeijer, A. Touw, M. van Herk, and J.V. Lebesque. Definition of the prostate in CT and MRI: a multi-observer study. *International Journal of Radiation Oncology, Biology, Physics*, 43(1):57–66, 1999.

[62] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[63] G.K. Rohde, A. Aldroubi, and B.M. Dawant. The adaptive bases algorithm for intensity-based nonrigid image registration. *IEEE Transactions on Medical Imaging*, 22(11):1470–1479, 2003.

[64] T. Rohlfing, R. Brandt, R. Menzel, and C.R. Maurer Jr. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. *NeuroImage*, 21(4):1428–1442, 2004.

[65] T. Rohlfing, C.R. Maurer Jr., D.A. Bluemke, and M.A. Jacobs. Volume-preserving nonrigid registration of MR breast images using free-form deformation with an incompressibility constraint. *IEEE Transactions on Medical Imaging*, 22(6):730–741, 2003.

[66] T. Rohlfing, D.B. Russakoff, and C.R. Maurer Jr. Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. *IEEE Transactions on Medical Imaging*, 23(8):983–994, 2004.

[67] G. Rudolph. Convergence rates of evolutionary algorithms for a class of convex objective functions. *Control and Cybernetics*, 26(3):375–390, 1997.

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

[69] J. Sacks. Asymptotic distribution of stochastic approximation procedures. *Annals of Mathematical Statistics*, 29(2):373–405, 1958.

[70] J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1):article 32, 2005.

[71] Z.-J. Shi and J. Shen. Step-size estimation for unconstrained optimization methods. *Computational & Applied Mathematics*, 24(3):399–416, 2005.

[72] R. Sitaram, N. Weiskopf, A. Caria, R. Veit, M. Erb, and N. Birbaumer. fMRI brain-computer interfaces. *IEEE Signal Processing Magazine*, 25(1):95–106, 2008.

[73] I. Sluimer, M. Prokop, and B. van Ginneken. Toward automated segmentation of the pathological lung in CT. *IEEE Transactions on Medical Imaging*, 24(8):1025–1038, 2005.

[74] J.C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992.

[75] J.C. Spall. Implementation of the simultaneous perturbation method for stochastic optimization. *IEEE Transactions on Aerospace and Electronic Systems*, 34(3):817–823, 1998.

[76] M. Staring, S. Klein, and J.P.W. Pluim. Nonrigid registration with tissue-dependent filtering of the deformation field. *Physics in Medicine and Biology*, 52(23):6879–6892, 2007.

[77] M. Staring, S. Klein, and J.P.W. Pluim. A rigidity penalty term for nonrigid registration. *Medical Physics*, 34(11):4098–4108, 2007.

[78] C. Studholme, C. Drapaca, B. Iordanova, and V. Cardenas. Deformation-based mapping of volume change from serial brain MRI in the presence of local tissue contrast change. *IEEE Transactions on Medical Imaging*, 25(5):626–639, 2006.

[79] C. Studholme, D.L.G. Hill, and D.J. Hawkes. Automated 3-D registration of MR and CT images of the head. *Medical Image Analysis*, 1(2):163–175, 1996.

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

[81] T.A. Sundaram and J.C. Gee. Towards a model of lung biomechanics: pulmonary kinematics via registration of serial lung images. *Medical Image Analysis*, 9(6):524–537, 2005.

[82] P. Thévenaz, M. Bierlaire, and M. Unser. Halton sampling for image registration based on mutual information. *Sampling Theory in Signal and Image Processing*, 7(2):141–171, 2008.

[83] P. Thévenaz, T. Blu, and M. Unser. Interpolation revisited. *IEEE Transactions on Medical Imaging*, 19(7):739–758, July 2000.

[84] P. Thévenaz, U.E. Ruttimann, and M. Unser. A pyramid approach to subpixel registration based on intensity. *IEEE Transactions on Image Processing*, 7(1):27–41, 1998.

[85] P. Thévenaz and M. Unser. Optimization of mutual information for multiresolution image registration. *IEEE Transactions on Image Processing*, 9(12):2083–2099, 2000.

[86] M. Unser. Splines: A perfect fit for signal and image processing. *IEEE Signal Processing Magazine*, 16(6):22–38, 1999.

[87] M. Unser, A. Aldroubi, and M. Eden. B-spline signal processing: Part I - Theory. *IEEE Transactions on Signal Processing*, 41(2):821–833, 1993.

[88] C.A. van Iersel, H.J. de Koning, G. Draisma, W.P.T.M. Mali, E.Th. Scholten, K. Nackaerts, M. Prokop, J.D.F. Habbema, M. Oudkerk, and R.J. van Klaveren. Risk-based selection from the general population in a screening trial: Selection criteria, recruitment and power for the Dutch-Belgian randomised lung cancer multi-slice CT screening trial (NELSON). *International Journal of Cancer*, 120(4):868–874, 2007.

[89] E.M. van Rikxoort, Y. Arzhaeva, and B. van Ginneken. A multi-atlas approach to automatic segmentation of the caudate nucleus in MR brain images. In T. Heimann, M. Styner, and B. van Ginneken, editors, *3D Segmentation In The Clinic: A Grand Challenge*, pages 29–36, 2007. `http://mbi.dkfz-heidelberg.de/grand-challenge2007`.

[90] G.M. Villeirs and G.O. De Meerleer. Magnetic resonance imaging (MRI) anatomy of the prostate and application of MRI in radiotherapy planning. *European Journal of Radiology*, 63(3):361–368, 2007.

[91] G.M. Villeirs, K. Van Vaerenbergh, L. Vakaet, S. Bral, F. Claus, W.J. De Neve, K.L. Verstraete, and G.O. De Meerleer. Interobserver delineation variation using CT versus combined CT+MRI in intensity-modulated radiotherapy for prostate cancer. *Strahlentherapie und Onkologie*, 181(7):424–430, 2005.

[92] P. Viola and W.M. Wells III. Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24(2):137–154, 1997.

[93] H.A. Vrooman, C.A. Cocosco, F. van der Lijn, R. Stokking, M.A. Ikram, M.W. Vernooij, M.M.B. Breteler, and W.J. Niessen. Multi-spectral brain tissue segmentation using automatically trained k-nearest-neighbor classification. *NeuroImage*, 37(1):71–81, 2007.

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

[95] J. West, J.M. Fitzpatrick, M.Y. Wang, B.M. Dawant, C.R. Maurer Jr., 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, G.Q. Maguire Jr., 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*, 21(4):554–566, 1997.

[96] G. Yin, G. Rudolph, and H.-P. Schwefel. Analyzing $(1, \lambda)$ evolution strategy via stochastic approximation methods. *Evolutionary Computation*, 3(4):473–489, 1995.

[97] Y. Zhu, S. Williams, and R. Zwiggelaar. Computer technology in detection and staging of prostate carcinoma: A review. *Medical Image Analysis*, 10:178–199, 2006.

[98] Y. Zhu, S. Williams, and R. Zwiggelaar. A hybrid ASM approach for sparse volumetric data segmentation. *Pattern Recognition and Image Analysis*, 17(2):252–258, 2007.

[99] B. Zitová and J. Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21(11):977–1000, 2003.

Samenvatting

Een belangrijk onderwerp in de medische beeldverwerking is de spatiële normalisatie van beelden. Het proces van spatiële normalisatie wordt *beeldregistratie* genoemd. De beelden kunnen afkomstig zijn van verschillende modaliteiten (MR, CT), van verschillende momenten in de tijd (bij vervolgstudies), en van verschillende patiënten (in het geval van populatiestudies). Beeldregistratie kan wiskundig geformuleerd worden als een optimalisatieprobleem (minimalisatie van een kostenfunctie). In dit proefschrift worden oplossingsmethoden onderzocht voor dit optimalisatieprobleem. Een samenvatting van elk hoofdstuk wordt hieronder gegeven:

Hoofdstuk 2 In de literatuur wordt een groot aantal beeldregistratiemethoden beschreven. Helaas is er niet één methode die in alle mogelijke toepassingsgebieden werkt. Dit hoofdstuk presenteert `elastix`: een vrij beschikbaar computerprogramma voor intensiteitsgebaseerde medische beeldregistratie. De programmatuur bestaat uit een verzameling algoritmes die vaak gebruikt worden voor de registratie van medische beelden. Door de modulaire opbouw van `elastix` kan de gebruiker voor een specifieke applicatie snel verschillende registratiemethoden configureren, testen en vergelijken. Het programma wordt aangestuurd door middel van tekstcommando's. Dit vergemakkelijkt de bewerking van grote hoeveelheden data met behulp van scripts. Het gebruik van `elastix` in vergelijkende studies is geïllustreerd met drie voorbeeldexperimenten, waarin individuele componenten van de registratiemethode zijn gevarieerd. De `elastix`-programmatuur is gebruikt voor alle experimenten die in dit proefschrift beschreven staan.

Hoofdstuk 3 Een vaak gebruikte techniek voor de niet-rigide registratie van medische beelden is gebaseerd op het maximaliseren van de hoeveelheid *wederzijdse informatie* tussen de beelden, in combinatie met een deformatieveld dat is geparametriseerd door kubische B-splines. De coördinatentransformatie die de twee afbeeldingen spatieel aan elkaar relateert wordt gevonden door middel van een iteratieve optimalisatieprocedure. Dit hoofdstuk vergelijkt de prestaties van acht optimalisatiemethoden: gradiëntdaling (met twee verschillende algoritmes voor het bepalen van de stapgrootte), quasi-Newton, niet-lineaire geconjugeerde gradiënt, Kiefer-Wolfowitz, simultane perturbatie, Robbins-Monro (RM) en een evolutiestrategie. Extra aandacht is besteed aan het reduceren van de rekentijd door niet alle voxels te gebruiken voor het berekenen van de kostenfunctie en haar afgeleiden. De optimalisatiemethoden zijn getest op kunstmatig gedeformeerde CT scans van het hart, CT scans van de thorax genomen op verschillende tijdstippen, en MR scans van de prostaat gemaakt volgens een BFFE, T1 en T2 protocol. De registratienauwkeurigheid is gemeten door de overlap te berekenen van gesegmenteerde randstructuren. Precisie en convergentiegedrag zijn bestudeerd door deformatievelden te vergelijken. De resultaten van de experimenten laten zien dat in de meeste applicaties de RM methode de beste keus is. Met deze techniek kan de rekentijd per iteratie ongeveer met een factor 500 verminderd worden, zonder dat de convergentiesnelheid aangetast wordt, door in elke iteratie de gradiënt van de kostenfunctie slechts op een klein deel van de voxels uit te rekenen. Deze voxels worden willekeurig gekozen, in elke iteratie opnieuw. Van de resterende methoden resulteren quasi-Newton en de methode van geconjugeerde gradiënten in een marginaal hogere precisie dan RM, ten koste van een veel hogere rekentijd.

Hoofdstuk 4 Dit hoofdstuk presenteert een stochastische optimalisatiemethode bedoeld voor intensiteitsgebaseerde registratie van monomodale beeldparen. De methode is gebaseerd op de RM stochastische gradiëntdalingsmethode en voegt daaraan een conditioneringsmatrix toe. De afleiding van de conditioneringsmatrix is gebaseerd op de observatie dat, na registratie, het gedeformeerde bewegende beeld er ongeveer hetzelfde moet uitzien als het niet-bewegende beeld. Met behulp van deze kennis vooraf kan de Hessiaan worden benaderd op het minimum van de registratiekostenfunctie, zonder dat de coördinatentransformatie die correspondeert met dit minimum bekend is. De methode is gevalideerd op 3D functionele MRI tijdreeksen en 3D CT scans van de thorax. De resultaten van de experimenten laten zien dat de geconditioneerde stochastische gradiëntdalingsmethode (PSGD) de convergentiesnelheid verbetert en de parameterselectie versimpelt, in vergelijking met Robbins-Monro.

Hoofdstuk 5 Dit hoofdstuk beschrijft een stochastische gradiëntdalingsoptimalisatiemethode met een adaptieve stapgrootte. De methode is gebaseerd op het theoretische werk van Plakhov en Cruz. Onze belangrijkste methodologische bijdrage is de afleiding van een beeldgebaseerde procedure voor het instellen van de belangrijkste vrije parameters. De procedure maakt gebruik van de karakteristieke vorm van de kostenfunctie in intensiteitsgebaseerde beeldregistratie. Ook wordt rekening gehouden met de theoretische convergentievoorwaarden van de optimalisatiemethode. Het resulterende adaptieve stochastische gradiëntdalingsalgoritme (ASGD) is vergeleken met een standaard, niet-adaptief RM algoritme. Zowel ASGD als RM gebruiken een stochastische bemonsteringstechniek om het optimalisatieproces te versnellen. Registratie-experimenten zijn gedaan op 3D CT en MR data van het hoofd, de longen en de prostaat, met verschillende similariteitsmaten en transformatiemodellen. De resultaten tonen aan dat ASGD robuust is voor deze variaties in het registratieraamwerk. Ook is ASGD minder gevoelig voor de instelling van parameters die door de gebruiker moeten worden opgegeven. De RM methode heeft als belangrijkste nadeel dat de stapgroottefunctie vooraf bepaald moet worden. De ASGD methode lost dit op.

Hoofdstuk 6 In dit hoofdstuk wordt een automatische methode gepresenteerd voor het segmenteren van de prostaat (inclusief de seminale vesikels) in 3D MR scans. De methode is gebaseerd op niet-rigide registratie van een set vooraf gelabelde atlasbeelden. Elk atlasbeeld wordt niet-rigide geregistreerd met het doelbeeld. Daarna worden de gedeformeerde atlaslabelbeelden gefuseerd om een segmentatie te verkrijgen van het patëntbeeld. De voorgestelde methode is geëvalueerd op 50 klinische scans, die elk handmatig gesegmenteerd zijn door drie experts. De Dice similariteitscoefficiënt (DSC) is gebruikt om de overlap tussen automatische en handmatige segmentaties te quantificeren. De invloed van een aantal factoren op de prestaties van de segmentatiemethode is onderzocht. Voor de registratie zijn twee similariteitsmaten vergeleken: wederzijdse informatie en een gelocaliseerde variant van de wederzijdse informatie. De laatste blijkt superieur te zijn ten opzichte van de eerste (mediaan ΔDSC ≈ 0.02, $p < 0.01$ volgens een gepaarde tweezijdige Wilcoxon test) en kost geen extra rekentijd dankzij het gebruik van een stochastisch optimalisatieschema gebaseerd op de onderliggende principes van de RM methode. Voor de fusie van de atlaslabelbeelden is een "meeste-stemmen-gelden" regel uitgeprobeerd en de zogenaamde "simultaneous truth and performance level estimation" (STAPLE) methode, beide met en zonder een voorafgaande atlasselectie. De verschillen tussen de genoemde fusiemethoden blijken minimaal te zijn en meestal niet statistisch significant ($p > 0.05$). Om een idee te krijgen van de invloed van de samenstelling van de atlas zijn twee atlassen vergeleken. De eerste set bestaat uit 38 scans van gezonde vrijwilligers. De tweede set is gegenereerd door steeds één van de 50 klinische scans te gebruiken voor evaluatie en de rest als atlas. De tweede atlas leverde substantieel betere resultaten op (ΔDSC $= 0.04$, $p < 0.01$), wat benadrukt hoe belangrijk het is om de atlas zorgvuldig samen te stellen. Met de beste instellingen bereiken we een mediale DSC van ongeveer 0.85. Dit komt dicht in de buurt van de DSC tussen de experts, die een mediaan van 0.87 bereiken. De segmentatienauwkeurigheid is vooral goed op de grens tussen prostaat en rectum, waar de fout beperkt blijft tot 1 mm in 50% van de gevallen en tot 1.5 mm in 75% van de gevallen.

In de introductie van dit proefschrift (Hoofdstuk 1) hebben we de volgende onderzoeksvragen geformuleerd:

- Welke optimalisatiemethode kan men het beste gebruiken voor registratie?

- Hoe kan men redelijke waarden bepalen voor de door de gebruiker te bepalen parameters van de desbetreffende optimalisatiemethode (als die er zijn)?

De experimentele resultaten uit dit proefschrift laten zien dat de RM stochastische gradiëntdalingsmethode, of een van haar verbeterde versies PSGD en ASGD, een goede keuze is in veel toepassingen. De stochastische bemonsteringstechniek die door deze methoden gebruikt wordt, zorgt ervoor dat de rekentijd per iteratie substantieel gereduceerd wordt, terwijl de convergentie behouden blijft. De basale RM methode heeft als nadeel dat de gebruiker een aantal belangrijk parameters, gerelateerd aan de stapgrootte in elke iteratie, handmatig moet instellen. De PSGD en ASGD zijn hier geïntroduceerd om dat probleem op te lossen. De PSGD methode is bedoeld voor monomodale registratieproblemen. De ASGD methode is algemener toepasbaar, maar doet een aantal simplificerende aannames die niet altijd terecht zouden kunnen zijn.

Natuurlijk zijn de experimenten niet uitputtend. Het is onmogelijk om te claimen dat de conclusies in alle situaties gelden. Echter, we hebben goede resultaten laten zien op een grote verscheidenheid aan registratieproblemen. Testen zijn gedaan met rigide, affiene en niet-rigide transformatiemodellen, met verschillende kostenfuncties (gemiddeld gekwadrateerd verschil, genormaliseerde correlatie, wederzijdse informatie, genormaliseerde wederzijdse informatie, locale wederzijdse informatie), verschillende beeldmodaliteiten en verscheidene anatomische structuren (hart, longen, prostaat, hersenen). De `elastix`-programmatuur beschreven in Hoofdstuk 2 maakt het mogelijk voor de lezer om de experimenten over te doen voor zijn/haar eigen toepassing.

Dit proefschrift is gericht op de registratie van *medische* beelden, zoals al gesuggereerd wordt door de titel. Het feit dat het gaat om medische beelden roept de volgende vraag op: "Wat is de klinische impact van deze studie?". De resultaten die gepresenteerd zijn in dit proefschrift tonen duidelijk aan dat de keuze van optimalisatiemethode een grote invloed kan hebben op de rekentijd, de robuustheid, en op de nauwkeurigheid. Dit zijn allemaal relevante aspecten bij gebruik in de kliniek. We hebben laten zien hoe men nauwkeurige niet-rigide registratie van grote 3D beelden kan realiseren binnen enkele minuten. De automatische prostaatintekeningsmethode die gepresenteerd is in Hoofdstuk 6 is een voorbeeld van een toepassing waar snelle registratie een vereiste is. Bijna 50 registraties zijn nodig voor de segmentatie van één enkele prostaat. Dit zou praktisch onmogelijk zijn met een registratie-algoritme dat uren aan rekentijd nodig heeft.

Door het vrij beschikbaar maken van de broncode van de algoritmes, als een deel van `elastix`, wordt het voor onderzoekers in de medische beeldverwerking mogelijk om de methoden direct in hun eigen applicaties te integreren. We hopen dat de algoritmes beschreven in dit proefschrift aldus hun weg zullen vinden naar de klinische praktijk.

Publications

Papers in international journals

- S. Klein, J.P.W. Pluim, M. Staring, and M.A. Viergever. Adaptive stochastic gradient descent optimisation for image registration. *International Journal of Computer Vision*, in press.

- S. Klein, U.A. van der Heide, I.M. Lips, M. van Vulpen, M. Staring, J.P.W. Pluim. Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information. *Medical Physics*, 35(4):1407–1417, 2008.

- S. Klein, M. Staring, J.P.W. Pluim. Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines. *IEEE Transactions on Image Processing*, 16(12):2879–2890, 2007.

- M. Staring, S. Klein, J.P.W. Pluim. A rigidity penalty term for nonrigid registration. *Medical Physics*, 34(11):4098–4108, 2007.

- M. Staring, S. Klein, J.P.W. Pluim. Nonrigid registration with tissue-dependent filtering of the deformation field. *Physics in Medicine and Biology*, 52(23):6879–6892, 2007.

- S. Klein, M. Staring, P. Andersson, and J.P.W. Pluim. Preconditioned stochastic gradient descent optimisation for monomodal image registration. *Submitted*.

- S. Klein, M. Staring, K. Murphy, M.A. Viergever, and J.P.W. Pluim. `elastix`: a toolbox for intensity-based medical image registration. *Submitted*.

- M. Staring, U.A. van der Heide, S. Klein, M.A. Viergever, and J.P.W. Pluim. Registration of cervical MRI using multifeature mutual information. *Submitted*.

- M. Staring, J.P.W. Pluim, B. de Hoop, S. Klein, B. van Ginneken, H. Gietema, G. Nossent, C. Schaefer-Prokop, S. van de Vorst, and M. Prokop. Influence of image subtraction on the assessment of volume and density change in ground-glass opacities in chest CT. *Submitted*.

- M.K. Chmarra, S. Klein, J.C.F. de Winter, F.-W. Jansen, and J. Dankelman. How to objectively classify surgeons based on their operative skills? *Submitted*.

Papers in conference proceedings

- K. Murphy, B. van Ginneken, J.P.W. Pluim, S. Klein, and M. Staring. Semi-Automatic Reference Standard Construction for Quantitative Evaluation of Lung CT Registration. In *Proceedings of MICCAI 2008*, in press.

- E.M. van Rikxoort, I. Išgum, M. Staring, S. Klein, B. van Ginneken. Adaptive local multi-atlas segmentation: application to heart segmentation in chest CT scans. In *Proceedings of SPIE Medical Imaging 2008*, in press.

- S. Klein, U.A. van der Heide, M. Staring, A.N.T.J. Kotte, B.W. Raaymakers, J.P.W. Pluim. Segmentation of the prostate in MR images by atlas matching using localised mutual information. In D.A. Jaffray, M. Sharpe, J. van Dyk, and J.P. Bissonnette, editors, *XVth International Conference on the use of Computers in Radiation Therapy*, volume 2, pages 585–589, 2007.

- S. Klein, U.A. van der Heide, B.W. Raaymakers, A.N.T.J. Kotte, M. Staring, J.P.W. Pluim. Segmentation of the prostate in MR images by atlas matching. In J.A. Fessler and T.S. Denney Jr., editors, *4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1300–1303, 2007.

- M. Staring, S. Klein, J.P.W. Pluim. Nonrigid registration using a rigidity constraint. In J.M. Reinhardt, J.P.W. Pluim, editors, *Medical Imaging: Image Processing*, volume 6144 of *Proceedings of SPIE*, pages 614413-1 – 614413-10, 2006.

- S. Klein, M. Staring, J.P.W. Pluim. A comparison of acceleration techniques for nonrigid medical image registration. In J.P.W. Pluim, B. Likar, F.A. Gerritsen, editors, *International Workshop on Biomedical Image Registration*, volume 4057 of *Lecture Notes in Computer Science*, pages 151–159, 2006.

- M. Staring, S. Klein, J.P.W. Pluim. Evaluation of a rigidity penalty term for nonrigid registration. In A. Bartoli, N. Navab, V. Lepetit, editors, *Workshop on Image Registration in Deformable Environments*, pages 41–50, 2006.

- M. Staring, S. Klein, J.P.W. Pluim. Nonrigid registration with adaptive, content-based filtering of the deformation field. In J.M. Fitzpatrick, J.M. Reinhardt, editors, *Medical Imaging: Image Processing*, volume 5747 of *Proceedings of SPIE*, pages 212–221, 2005.

- S. Klein, M. Staring, J.P.W. Pluim. Comparison of gradient approximation techniques for optimisation of mutual information in nonrigid registration. In J.M. Fitzpatrick, J.M. Reinhardt, editors, *Medical Imaging: Image Processing*, volume 5747 of *Proceedings of SPIE*, pages 192–203, 2005.

- S. Klein, A.M. Bazen, R.N.J. Veldhuis. Fingerprint image segmentation based on hidden Markov models. In *Proceedings of ProRISC 2002, 13th Annual Workshop on Circuits, Systems, and Signal Processing*, 2002.

Curriculum Vitæ

Stefan Klein was born in Almelo, the Netherlands, on April 20, 1978. From 1996 to 2002, he studied Mechanical Engineering at the University of Twente, Netherlands. His MSc project was on the segmentation of fingerprint images using Hidden Markov Models. From 2003 to 2008 he worked as a PhD student at the Image Sciences Institute (ISI), University Medical Center Utrecht, Utrecht, the Netherlands. This thesis is the result of that period. Since 2008, he is with the Biomedical Imaging Group Rotterdam (BIGR), Erasmus MC, Rotterdam, the Netherlands.