Evaluation of Optimization Methods for Nonrigid Medical Image Registration Using Mutual Information and B-Splines

Stefan Klein, Marius Staring, and Josien P. W. Pluim

***Abstract*—A popular technique for nonrigid registration of medical images is based on the maximization of their mutual information, in combination with a deformation field parameterized by cubic B-splines. The coordinate mapping that relates the two images is found using an iterative optimization procedure. This work compares the performance of eight optimization 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 optimization 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.**

***Index Terms*—B-splines, mutual information, nonrigid image registration, optimization, subsampling.**

I. 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 3-D images, ask for faster registration algorithms [1]. Other applications, such as brain shift estimation based on intraoperatively acquired ultrasound [2], 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 online updating of the treatment plan [3].

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 maximization of the *mutual information* similarity measure, in combination with a deformation field parameterized by *cubic B-splines* [4], [5]. The approach can be formulated as a minimization problem

$$\hat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu}} \mathcal{C}(\boldsymbol{\mu}; I_F, I_M) \tag{1}$$

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 optimization algorithm and on the initial alignment of the images. A regularization term $\mathcal{R}$ can be added to the cost function, to penalize undesirable deformations, and, consequently, to reduce the number of local minima

$$\mathcal{C}(\boldsymbol{\mu}; I_F, I_M) = -MI(\boldsymbol{\mu}; I_F, I_M) + \omega\mathcal{R}(\boldsymbol{\mu}). \tag{2}$$

In this equation, ω serves as a weighting factor for the regularization term. Well-known examples for $\mathcal{R}$ are the curvature term [6], the elastic energy [7], and the volume preserving penalty term [8]. With nonparametric registration techniques, which do not employ a parametric model of the deformation, a proper regularization term is essential to ensure smoothness (differentiability) of the deformation field [6]. In the parametric approach that we focus on, the regularization term may be superfluous, since the cubic B-spline basis functions are inherently smooth. However, additional regularization 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 optimization strategy is employed

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k + a_k \boldsymbol{d}_k, \quad k = 0, 1, 2, \ldots \tag{3}$$

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 optimization methods can be found in the literature [9]–[12], differing in the way a_k and $\boldsymbol{d}_k$ are computed. In

Manuscript received April 25, 2007; revised August 8, 2007. This work was supported by the Netherlands Organization for Scientific Research (NWO). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Scott T. Acton.

The authors are with the University Medical Center Utrecht, Image Sciences Institute, 3508GA Utrecht, The Netherlands (e-mail: stefan@isi.uu.nl; marius@isi.uu.nl; josien@isi.uu.nl).

Digital Object Identifier 10.1109/TIP.2007.909412

contrast to the field of *rigid* registration [13], no extensive comparison of optimization procedures has been done for *nonrigid* image registration problems.

In this paper, several optimization methods are compared with respect to speed, accuracy, precision, and robustness. The following methods are included in the study: gradient descent [9], quasi-Newton [14], nonlinear conjugate gradient [15], Kiefer–Wolfowitz [16], simultaneous perturbation [17], Robbins–Monro [18], and evolution strategy [19]. 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 III explains the optimization 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 optimization methods, because they expect exact derivatives. However, stochastic optimization methods may be able to deal with it. The technique has been proposed for rigid registration problems [20], but its effect on *nonrigid* registration has not been evaluated in the literature. Section IV discusses the topic more extensively.

The experiments and results are described in Section V. The optimization 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 Section VI.

II. 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.* [5], Mattes *et al.* [4], and Thévenaz and Unser [21].

The registration method uses cubic B-splines to parameterize the deformation field, and mutual information as the similarity measure. Several implementations for the computation of mutual information can be found in the literature [20]–[23]. The approach described by Thévenaz and Unser [21] is used here. The mutual information is defined as follows:

$$MI(\boldsymbol{\mu}; I_F, I_M) = \sum_{m\in L_M}\sum_{f\in L_F} p(f,m;\boldsymbol{\mu}) \times \log_2\left(\frac{p(f,m;\boldsymbol{\mu})}{p_F(f)p_M(m;\boldsymbol{\mu})}\right) \quad (4)$$

where L_F and L_M are sets of regularly spaced intensity bin centers, 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

$$p(f,m;\boldsymbol{\mu}) = \frac{1}{|I_F|}\sum_{\boldsymbol{x}_i\in I_F} w_F\left(f/\sigma_F - I_F(\boldsymbol{x}_i)/\sigma_F\right) \times w_M\left(m/\sigma_M - I_M\left(\boldsymbol{x}_i + \boldsymbol{u}_{\boldsymbol{\mu}}(\boldsymbol{x}_i)\right)/\sigma_M\right) \quad (5)$$

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 $\partial MI/\partial\boldsymbol{\mu}$; see [4] and [21]. For w_F, a zeroth-order B-spline can be used [4]. 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 gray-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 paper have been performed with and without the regularization term $\mathcal{R}$. A regularization term is used that penalizes second-order derivatives of the deformation field

$$\mathcal{R} = \frac{1}{|I_F|}\sum_{\boldsymbol{x}_i\in I_F}\sum_{p,q,r}\left(\frac{\partial^2 u_p}{\partial x_q \partial x_r}(\boldsymbol{x}_i)\right)^2. \quad (6)$$

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

To guide the optimization towards the desired local minimum of the cost function, multiresolution strategies are often employed. For extensive overviews on this subject, we refer to [24] and [25]. 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.

III. Optimization Methods

All optimization algorithms studied in this paper can be written in the form of (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 minimize the cost function $\mathcal{C}$ along the search direction $\boldsymbol{d}_k$

$$a_k = \arg\min_a \mathcal{C}(\boldsymbol{\mu}_k + a\boldsymbol{d}_k). \quad (7)$$

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 (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 optimization 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 Section II, an analytic expression for the derivative of the mutual information is available. Some optimization methods require exact evaluation of this expression. Other methods are satisfied with an approximation.

A. Gradient Descent (GDD and GDL)

The gradient descent method [9] takes steps in the direction of the negative gradient of the cost function

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - a_k \boldsymbol{g}(\boldsymbol{\mu}_k) \tag{8}$$

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

In this paper, 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 Section III-D), 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 Sections III-B and C. 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 [10] that there exist a $K \geq 0$ and $\rho > 0$, such that the following expression, holds:

$$\frac{\|\boldsymbol{\mu}_{k+1} - \hat{\boldsymbol{\mu}}\|}{\|\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}}\|} \leq \rho, \quad \text{for all } k \geq K. \tag{9}$$

This means that the method has a linear rate of convergence. If $\rho \geq 1$, the term "sublinear convergence" is used [26].

B. Quasi-Newton (QN)

QN methods [9], [14] are inspired by the well-known Newton–Raphson algorithm, which is given by

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - [H(\boldsymbol{\mu}_k)]^{-1}\, \boldsymbol{g}(\boldsymbol{\mu}_k) \tag{10}$$

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 optimization problems such as nonrigid registration. QN 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. QN 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:

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - a_k L_k \boldsymbol{g}(\boldsymbol{\mu}_k). \tag{11}$$

Given certain conditions, many QN methods can be shown to be *superlinearly* convergent [14]

$$\lim_{k\to\infty} \frac{\|\boldsymbol{\mu}_{k+1} - \hat{\boldsymbol{\mu}}\|}{\|\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}}\|} \to 0. \tag{12}$$

Many ways to construct the series $\{L_k\}$ are proposed in the literature [9], [14], 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 [9]. It uses the following update rule for L_k:

$$L_{k+1} = \left(I - \frac{\boldsymbol{s}\boldsymbol{y}^T}{\boldsymbol{s}^T\boldsymbol{y}}\right) L_k \left(I - \frac{\boldsymbol{y}\boldsymbol{s}^T}{\boldsymbol{s}^T\boldsymbol{y}}\right) + \frac{\boldsymbol{s}\boldsymbol{s}^T}{\boldsymbol{s}^T\boldsymbol{y}} \tag{13}$$

where 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) [27], which eliminates the need for storing the matrix L_k in memory.

Following the implementation described in [27], we use the inexact line search routine described by Moré and Thuente [28]. It determines a_k such that the so-called *strong Wolfe conditions* are satisfied

$$\mathcal{C}(\boldsymbol{\mu}_{k+1}) \leq \mathcal{C}(\boldsymbol{\mu}_k) + c_1 a_k \boldsymbol{d}_k^T \boldsymbol{g}(\boldsymbol{\mu}_k) \tag{14}$$

$$\left|\boldsymbol{d}_k^T \boldsymbol{g}(\boldsymbol{\mu}_{k+1})\right| \leq c_2 \left|\boldsymbol{d}_k^T \boldsymbol{g}(\boldsymbol{\mu}_k)\right| \tag{15}$$

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 optimization algorithm [see (3)], which equals $-L_k \boldsymbol{g}(\boldsymbol{\mu}_k)$ in the case of QN methods. The first Wolfe condition (14) demands a sufficient decrease of the cost function value. The second Wolfe condition (15) enforces reasonable progress towards a stationary point of the cost function, where the derivative vanishes. For optimization 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 [29]. To realize superlinear convergence it is important to always try a gain factor $a_k = 1$ first [9]. 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 optimization is assumed to have converged.

C. Nonlinear Conjugate Gradient (NCG)

The development of conjugate gradient methods started with the *linear* conjugate gradient method [30]. This routine was designed for solving a system of linear equations, which is equiv-

alent to the minimization of a quadratic cost function. The *nonlinear* conjugate gradient method is an extension suitable for minimizing general nonlinear functions [9], [15]. The NCG algorithm follows the general iterative scheme (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 $\boldsymbol{d}_{k-1}$

$$\boldsymbol{d}_k = -\boldsymbol{g}(\boldsymbol{\mu}_k) + \beta_k \boldsymbol{d}_{k-1}. \tag{16}$$

Several expressions for the scalar β_k have been proposed in the literature [15], including

$$\text{Dai} - \text{Yuan}: \quad \beta_k^{\text{DY}} = \frac{\boldsymbol{g}_k^T \boldsymbol{g}_k}{\boldsymbol{d}_{k-1}^T (\boldsymbol{g}_k - \boldsymbol{g}_{k-1})} \tag{17}$$

$$\text{Hestenes} - \text{Stiefel}: \quad \beta_k^{\text{HS}} = \frac{\boldsymbol{g}_k^T (\boldsymbol{g}_k - \boldsymbol{g}_{k-1})}{\boldsymbol{d}_{k-1}^T (\boldsymbol{g}_k - \boldsymbol{g}_{k-1})} \tag{18}$$

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 [15]. In our study, we use a hybrid version, proposed in [31] and shown to be very efficient compared to other methods

$$\beta_k = \max\left(0, \min\left(\beta_k^{\text{HS}}, \beta_k^{\text{DY}}\right)\right). \tag{19}$$

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 [32], 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 Section III-A), the same approach is used.

D. Stochastic Gradient Descent

The stochastic gradient descent method [12] follows the same scheme as the *deterministic* gradient descent, see (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:

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - a_k \widetilde{\boldsymbol{g}}_k. \tag{20}$$

Convergence to the solution $\hat{\boldsymbol{\mu}}$ can only be guaranteed [11] if the bias of the approximation error goes to zero

$$\text{E}(\widetilde{\boldsymbol{g}}_k) \to \boldsymbol{g}(\boldsymbol{\mu}_k), \quad \text{as } k \to \infty \tag{21}$$

where $\text{E}(\cdot)$ 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: KW, SP, and RM.

- Kiefer–Wolfowitz (KW): This method, originally proposed in [16], is based on a finite difference approximation of the derivative, given by

$$[\widetilde{\boldsymbol{g}}_k]_i = \frac{\mathcal{C}(\boldsymbol{\mu}_k + c_k \boldsymbol{e}_i) - \mathcal{C}(\boldsymbol{\mu}_k - c_k \boldsymbol{e}_i)}{2c_k} \tag{22}$$

  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 KW method assumes that only approximations of the cost function values are available

$$\widetilde{\mathcal{C}}_{ki}^{+} = \mathcal{C}(\boldsymbol{\mu}_k + c_k \boldsymbol{e}_i) + \varepsilon_{ki}^{+}$$

and

$$\widetilde{\mathcal{C}}_{ki}^{-} = \mathcal{C}(\boldsymbol{\mu}_k - c_k \boldsymbol{e}_i) + \varepsilon_{ki}^{-} \tag{23}$$

  where ε_{ki}^{+} and ε_{ki}^{-} represent the approximation errors. Substituting this in (22) yields the KW algorithm

$$[\widetilde{\boldsymbol{g}}_k]_i = \frac{\widetilde{\mathcal{C}}_{ki}^{+} - \widetilde{\mathcal{C}}_{ki}^{-}}{2c_k}. \tag{24}$$

  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 (23). For c_k, the following expression is commonly used:

$$c_k = c/(k+1)^{\gamma} \tag{25}$$

  where $c > 0$ and $0 \leq \gamma \leq 1$ are user-defined constants. Note that, for an N-dimensional optimization 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 [17], 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 two evaluations, independent of N

$$\widetilde{\mathcal{C}}_k^{+} = \mathcal{C}(\boldsymbol{\mu}_k + c_k \boldsymbol{\Delta}_k) + \varepsilon_k^{+}$$

and

$$\widetilde{\mathcal{C}}_k^{-} = \mathcal{C}(\boldsymbol{\mu}_k - c_k \boldsymbol{\Delta}_k) + \varepsilon_k^{-}. \tag{26}$$

  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

$$[\widetilde{\boldsymbol{g}}_k]_i = \frac{\widetilde{\mathcal{C}}_k^{+} - \widetilde{\mathcal{C}}_k^{-}}{2c_k [\boldsymbol{\Delta}_k]_i}. \tag{27}$$

The scalar c_k is defined according to (25). The simultaneous perturbation method has been used for rigid registration [33], but its performance has not been compared to other optimization methods.

- Robbins–Monro (RM): Whereas KW and SP construct a derivative estimate based on approximate evaluations of the cost function, the RM algorithm [18] does not specify how the derivative is computed. It assumes that an approximation of the derivative of the cost function is available

$$\widetilde{\boldsymbol{g}}_k = \boldsymbol{g}(\boldsymbol{\mu}_k) + \boldsymbol{\varepsilon}_k. \tag{28}$$

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 Section III-A.

In Section IV, 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 [34]:

$$a_k = a/(k+A)^\alpha \tag{29}$$

with $A \geq 1$. This will be used in our experiments. The same gain sequence is used by the GDD method (Section III-A). Theoretically optimal values for α are derived in [11] and [17]. For SP specifically, practical guidance for choosing a, A, and α is provided in [34]. 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 (9). Instead, the theoretical convergence properties are given in terms of the "asymptotic normality" of $(\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}})$

$$(k+1)^\beta(\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}}) \sim \mathcal{N}(\boldsymbol{m}, \Sigma), \quad \text{as } k \to \infty, \tag{30}$$

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 [11], [17], [35], and [36].

E. 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 [37]. The covariance matrix adaptation (CMA) ES [19] is generally considered to be the current state-of-the-art ES algorithm [38] 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 $\mathcal{N}$

$$\boldsymbol{d}_k^{(\ell)} \sim \mathcal{N}(\boldsymbol{0}, C_k), \quad \text{for } \ell = 1, 2, \ldots, \lambda. \tag{31}$$

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

$$\boldsymbol{d}_k = \sum_{p=1}^{P} w_p \boldsymbol{d}_k^{(p;\lambda)}. \tag{32}$$

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).

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 [19] 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 [19] 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 [38]–[40], for example. Experimental results with synthetic cost functions [41], [42] indicate that approximate (noisy) cost function evaluations can be dealt with to some degree.

IV. 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 ($t_\mathcal{C}$) and its derivative ($t_{\boldsymbol{g}}$) 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

$$t_\mathcal{C} \sim p|I_F| + q \tag{33}$$

$$t_{\boldsymbol{g}} \sim r|I_F| + sN \tag{34}$$

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 optimization 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 optimization process, a bias in the approximation error is avoided. This technique, which we call "stochastic subsampling," has been proposed before for rigid registration problems [20], 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 optimization 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 [4], [43]. 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 optimization method's name, for example QN-2. A downsampling factor of 1 corresponds to using the full image.

V. Experiments and Results

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

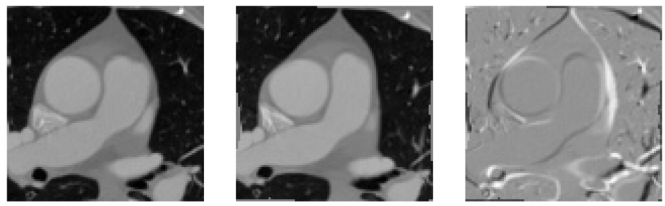
$$D(\boldsymbol{u}_1, \boldsymbol{u}_2) = \frac{1}{|I_F|} \sum_{\boldsymbol{x}_i \in I_F} \|\boldsymbol{u}_1(\boldsymbol{x}_i) - \boldsymbol{u}_2(\boldsymbol{x}_i)\|. \qquad (35)$$

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

$$\text{overlap} = \frac{2|V_1 \cap V_2|}{|V_1| + |V_2|} \cdot 100\%. \qquad (36)$$

This measure is known as the Dice similarity index [45]. 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 [46]. 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 $3 \times 3 \times 3$ kernel.



(a) (b) (c)

Fig. 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, and (c) the difference between (a) and (b). Voxels in the deformed volume that map outside the original image were set to 0.

The computational efficiency of the optimization method depends on the number of required iterations and the computation time per iteration. The computation 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 (34) can be neglected. Consequently, for GD, QN, NCG, and RM, the computation time per iteration equals $\eta t_{\boldsymbol{g}}$, with η the fraction of the voxels used to compute the derivative, and $t_{\boldsymbol{g}}$ the time required to compute the derivative using *all* voxels. All timing results in this paper are reported as a factor times $t_{\boldsymbol{g}}$. For example, for 512 iterations of QN-4 with 3-D images, we report a computation time of $512(1/4)^3 t_{\boldsymbol{g}} = 8t_{\boldsymbol{g}}$. Note that the computation times per iteration of KW, SP, and ES are not obviously related to $t_{\boldsymbol{g}}$. To express them as a factor times $t_{\boldsymbol{g}}$, we rely on experimental measurements. For each application, we also report the value of $t_{\boldsymbol{g}}$ in seconds (measured with an AMD Opteron 244, 1.8 GHz), to give an indication of the typical computation times.

A. 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 $\boldsymbol{u}_{\boldsymbol{\mu}}$ that is updated during optimization. The average displacement distance $D(\boldsymbol{u}_{\boldsymbol{\mu}}, \tilde{\boldsymbol{u}})$ can be used to assess the registration quality.

The registrations were performed on four 3-D 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 $97 \times 97 \times 97$ 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. Fig. 1 shows an example slice, its deformed version, and the difference image visualizing the initial misalignment.

The registrations were performed using a $10 \times 10 \times 10$ grid of B-spline control points to parameterize the deformation field

$\boldsymbol{u}_{\boldsymbol{\mu}}$, yielding $N = 3000$ parameters to be optimized. 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 regularization term was used, either. The maximum number of iterations was limited to 2000. Three constants must be set for the gain sequence (29) employed by the optimization 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 (25)]: $c = 1.0$ and $\gamma = 0.101$. The choices for α and γ are based on the recommendations in [34]. For ES, the initial step size a_0 was set to 1.0, and, following the recommendations in [19], the values $\lambda = 28$ and $P = 14$ were used. The stochastic optimization 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.

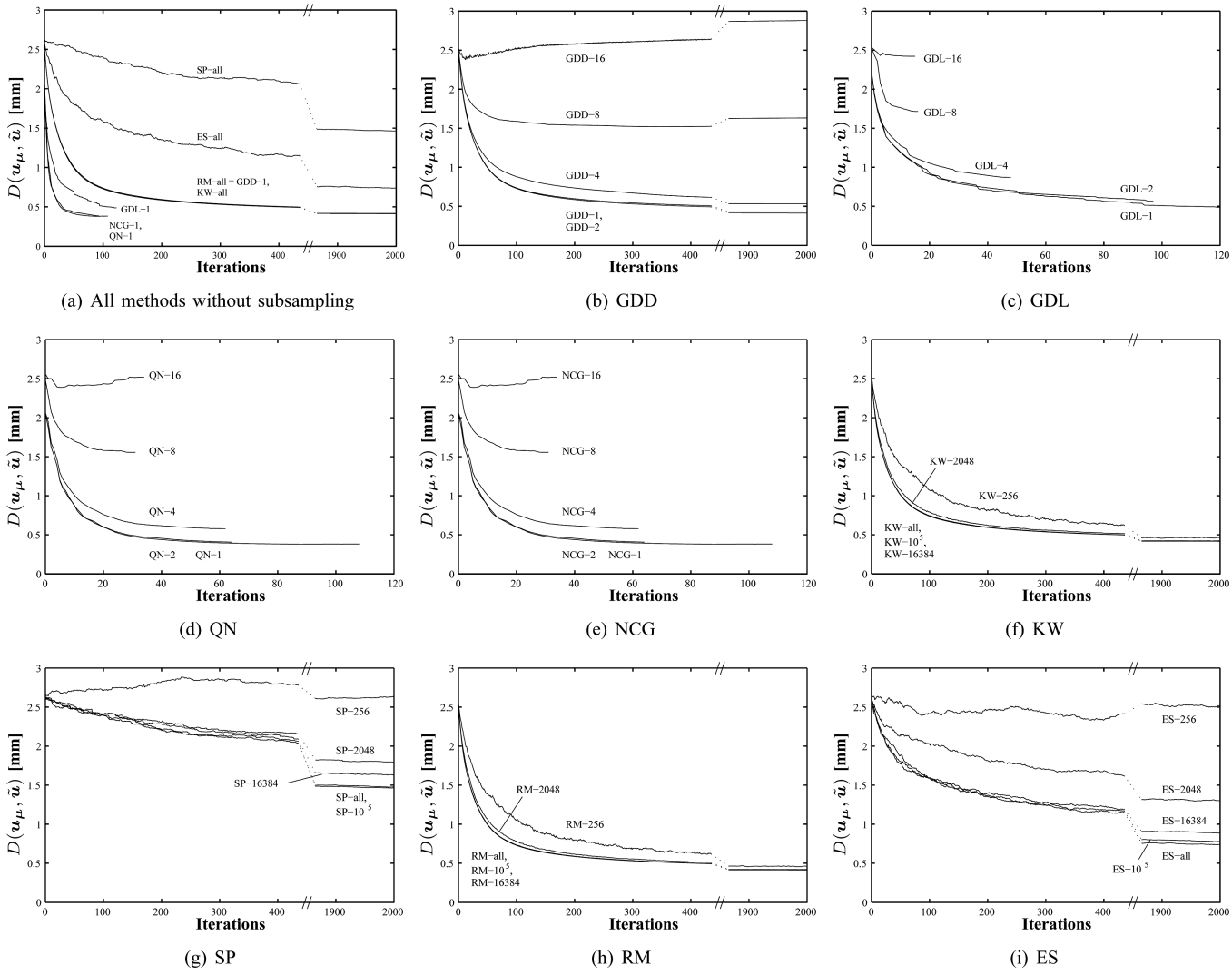
In this paper, we present the results for one of the four CT images. The outcome for the other images was similar. In Fig. 2(a), the convergence results are given for all methods, without subsampling. The error measure $D(\boldsymbol{u}_{\boldsymbol{\mu}_k}, \tilde{\boldsymbol{u}})$ is plotted against the number 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 Section III-B). 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 optimization method is presented in Fig. 2(b)–(i). Fig. 2(b)–(e) shows the effect of deterministic subsampling: downsampling by a factor of 4 or more degraded the registration results of GDD, GDL, QN, and NCG. Fig. 2(f)–(i) shows 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 $\eta t_{\boldsymbol{g}}$, with η the fraction of voxels used. One $t_{\boldsymbol{g}}$ was measured to be 20 s approximately. For KW, SP, and ES, the computation times needed for the cost function evaluations in each iteration were measured to be around $10\eta t_{\boldsymbol{g}}$, $0.9\eta t_{\boldsymbol{g}}$ $(= 2\eta t_{\mathcal{C}})$, and $13\eta t_{\boldsymbol{g}}$ $(= 28\eta t_{\mathcal{C}})$, 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, 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 visualize the acceleration realized 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 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 $\tilde{\boldsymbol{u}}$ 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 $\tilde{\boldsymbol{u}}$ 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. 4. 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 optimization 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, $\mathcal{C}(\boldsymbol{\mu}_k) < \mathcal{C}(\boldsymbol{\mu}_{k-1})$. Apparently, the lower cost function did not translate into a better accuracy. We have repeated the experiments using the regularization term $\mathcal{R}$ with $\omega = 500$ [see (2)]. This resolved the issue, but did not change the efficiency differences between the methods. The effect of regularization is studied further in the following sections.

B. 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 paper, we used data of five patients.



(a) All methods without subsampling (b) GDD (c) GDL

(d) QN (e) NCG (f) KW

(g) SP (h) RM (i) ES

Fig. 2. Convergence results for all methods. 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. (a) All methods without subsampling; (b) GDD; (c) GDL; (d) QN; (e) NCG; (f) KW; (g) SP; (h) RM; (i) ES.

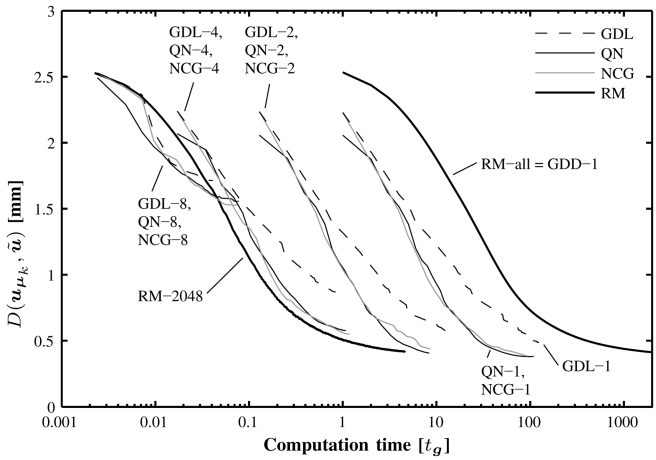


Fig. 3. Timing results for GDL, QN, NCG, and RM ($t_g \approx 20$ s).

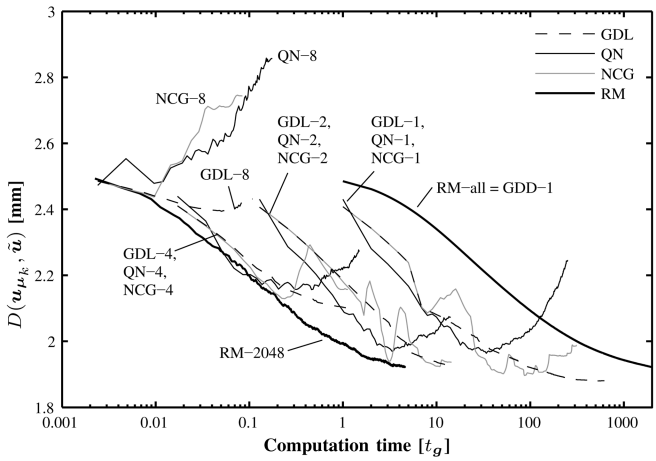


Fig. 4. Timing results on a more difficult registration problem ($t_g \approx 20$ s).

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

TABLE I
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 avg [t_g]	overlap avg ± sd [%]	precision avg ± sd [mm]	effect $\mathcal{R}$ avg ± sd [mm]	time avg [t_g]	overlap avg ± sd [%]	precision avg ± sd [mm]	effect $\mathcal{R}$ avg ± sd [mm]	overlap* avg ± sd [%]	precision* avg ± sd [mm]	time avg [t_g]	precision avg ± sd [mm]	effect $\mathcal{R}$ avg ± sd [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

B-spline control point spacing was set to 22 mm, yielding a grid of about $19 \times 19 \times 19$ control points; approximately 20000 parameters to optimize. 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 regularization term $\mathcal{R}$. For the weighting factor ω, a value of 500 was used. Without a regularization 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 regularization 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.* [1], [47]. 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 $7 \times 7 \times 7$ structuring element.

The results are located in the left part of Table I. Overlap and precision values were calculated after rigid registration and nonrigid registration using GDL, QN, NCG, and RM, all with regularization. The results for the five patients are summarized 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 s approximately. The time needed to calculate the derivative of the regularization term was not counted, since it could be implemented as a cascade of fast filter operations on the B-spline coefficients [48]. The fourth column ("effect $\mathcal{R}$") shows the average displacement distance between the solutions obtained with and without $\mathcal{R}$, indicating how the regularization 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 optimized 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 regularization term.

C. Motion Between Interfraction MR Prostate Scans

Prostate cancer treatment by radiation therapy requires an accurate localization 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 [3], [49]. 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 $512 \times 512 \times 90$ voxels, with a voxel size of $0.49 \times 0.49 \times 1.0$ mm. The T1 and T2 have a dimension of $256 \times 256 \times 25$ voxels, with highly anisotropic voxels of $0.8 \times 0.8 \times 4.0$ 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 optimized. For the T1-T2 experiments a grid resolution of $30 \times 30 \times 70$ 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 regularization 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 registration, 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 optimization 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 center and right part of Table I 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, $t_{\boldsymbol{g}}$ was around 56 s, for the T1-T2 registrations $t_{\boldsymbol{g}}$ was around 9 s.

VI. Conclusion

We have compared eight optimization methods for nonrigid registration based on the maximization of mutual information, in combination with a deformation field parameterized 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 optimization 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 regularization 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 generalized to the branch of nonparametric registration algorithms [6], [7].

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 [34], which work satisfactorily for the Robbins–Monro method, as well, in our experience. Note that in all experiments described in this paper 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. Using the Robbins–Monro approach, acceleration factors of approximately 500, compared to a basic gradient descent method, can be easily achieved on many 3-D nonrigid registration problems.

ACKNOWLEDGMENT

The authors would like to thank E. Kerkhof for providing the prostate segmentations. Additionally, this work benefited from the use of the Insight Segmentation and Registration Toolkit (ITK), an open source software developed as an initiative of the U.S. National Library of Medicine and available at http://www.itk.org.

REFERENCES

[1] I. Sluimer, M. Prokop, and B. van Ginneken, "Toward automated segmentation of the pathological lung in CT," *IEEE Trans. Med. Imag.*, vol. 24, no. 8, pp. 1025–1038, Aug. 2005.

[2] X. Pennec, P. Cachier, and N. Ayache, "Tracking brain deformations in time sequences of 3D US images," *Pattern Recognit. Lett.*, vol. 24, no. 4–5, pp. 801–813, 2003.

[3] B. Fei, J. L. Duerk, D. B. Sodee, and D. L. Wilson, "Semiautomatic nonrigid registration for the prostate and pelvic MR volumes," *Acad. Radiol.*, vol. 12, no. 7, pp. 815–824, 2005.

[4] 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 Trans. Med. Imag.*, vol. 22, no. 1, pp. 120–128, Jan. 2003.

[5] 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 Trans. Med. Imag.*, vol. 18, no. 8, pp. 712–721, Aug. 1999.

[6] B. Fischer and J. Modersitzki, "A unified approach to fast image registration and a new curvature based registration technique," *Linear Algebra Appl.*, vol. 380, pp. 107–124, 2004.

[7] G. E. Christensen and H. J. Johnson, "Consistent image registration," *IEEE Trans. Med. Imag.*, vol. 20, no. 7, pp. 568–582, Jul. 2001.

[8] 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 Trans. Med. Imag.*, vol. 22, no. 6, pp. 730–741, Jun. 2003.

[9] J. Nocedal and S. J. Wright, *Numerical Optimization*. New York: Springer-Verlag, 1999.

[10] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA: Athena Scientific, 1999.

[11] H. J. Kushner and D. S. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. New York: Springer-Verlag, 1978.

[12] H. J. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed. New York: Springer-Verlag, 2003.

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

[14] J. E. Dennis, Jr. and J. J. Moré, "Quasi-Newton methods, motivation and theory," *SIAM Rev.*, vol. 19, no. 1, pp. 46–89, 1977.

[15] Y.-H. Dai, "A family of hybrid conjugate gradient methods for unconstrained optimization," *Math. Comput.*, vol. 72, no. 243, pp. 1317–1328, 2003.

[16] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *Ann. Math. Statist.*, vol. 23, no. 3, pp. 462–466, 1952.

[17] J. C. Spall, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Trans. Autom. Control*, vol. 37, no. 3, pp. 332–341, 1992.

[18] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, 1951.

[19] N. Hansen and A. Ostermeier, "Completely derandomized self-adaptation in evolution strategies," *Evol. Comput.*, vol. 9, no. 2, pp. 159–195, 2001.

[20] P. Viola and W. M. Wells, III, "Alignment by maximization of mutual information," *Int. J. Comput. Vis.*, vol. 24, no. 2, pp. 137–154, 1997.

[21] P. Thévenaz and M. Unser, "Optimization of mutual information for multiresolution image registration," *IEEE Trans. Image Process.*, vol. 9, no. 12, pp. 2083–2099, Dec. 2000.

[22] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Trans. Med. Imag.*, vol. 16, no. 2, pp. 187–198, Feb. 1997.

[23] C. Studholme, D. L. G. Hill, and D. J. Hawkes, "Automated 3-D registration of MR and CT images of the head," *Med. Image Anal.*, vol. 1, no. 2, pp. 163–175, 1996.

[24] H. Lester and S. R. Arridge, "A survey of hierarchical non-linear medical image registration," *Pattern Recognit.*, vol. 32, no. 1, pp. 129–149, 1999.

[25] M. Lefébure and L. D. Cohen, "Image registration, optical flow and local rigidity," *J. Math. Imag. Vis.*, vol. 14, no. 2, pp. 131–147, 2001.

[26] L. O. Jay, "A note on Q-order of convergence," *BIT Numer. Math.*, vol. 41, no. 2, pp. 422–429, 2001.

[27] J. Nocedal, "Updating quasi-Newton matrices with limited storage," *Math. Comput.*, vol. 35, no. 151, pp. 773–782, 1980.

[28] J. J. Moré and D. J. Thuente, "Line search algorithms with guaranteed sufficient decrease," *ACM Trans. Math. Software*, vol. 20, no. 3, pp. 286–307, 1994.

[29] F. A. Potra and Y. Shi, "Efficient line search algorithm for unconstrained optimization," *J. Optim. Theor. Appl.*, vol. 85, no. 3, pp. 677–704, 1995.

[30] M. R. Hestenes and E. Stiefel, "Methods of conjugate gradients for solving linear systems," *J. Res. Nat. Bur. Stand.*, vol. 49, pp. 409–436, 1952.

[31] Y.-H. Dai, "An efficient hybrid conjugate gradient method for unconstrained optimization," *Ann. Oper. Res.*, vol. 103, pp. 33–47, 2001.

[32] H. Mukai, "Readily implementable conjugate gradient methods," *Math. Program.*, vol. 17, pp. 298–319, 1979.

[33] 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 Trans. Image Process.*, vol. 12, no. 12, pp. 1495–1511, Dec. 2003.

[34] J. C. Spall, "Implementation of the simultaneous perturbation method for stochastic optimization," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 34, no. 3, pp. 817–823, Jul. 1998.

[35] V. Fabian, "On asymptotic normality in stochastic approximation," *Ann. Math. Statist.*, vol. 39, no. 4, pp. 1327–1332, 1968.

[36] J. Sacks, "Asymptotic distribution of stochastic approximation procedures," *Ann. Math. Statist.*, vol. 29, no. 2, pp. 373–405, 1958.

[37] H.-G. Beyer and H.-P. Schwefel, "Evolution strategies—A comprehensive introduction," *Nat. Comput.*, vol. 1, no. 12, pp. 3–52, 2002.

[38] A. Auger, "Convergence results for the $(1, \lambda)$-SA-ES using the theory of ϕ-irreducible markov chains," *Theoret. Comput. Sci.*, vol. 334, pp. 35–69, 2005.

[39] G. Yin, G. Rudolph, and H.-P. Schwefel, "Analyzing $(1, \lambda)$ evolution strategy via stochastic approximation methods," *Evol. Comput.*, vol. 3, no. 4, pp. 473–489, 1995.

[40] G. Rudolph, "Convergence rates of evolutionary algorithms for a class of convex objective functions," *Contr. Cybern.*, vol. 26, no. 3, pp. 375–390, 1997.

[41] D. V. Arnold and H.-G. Beyer, "Evolution strategies with cumulative step length adaptation on the noisy parabolic ridge," *Nat. Comput.*, to be published.

[42] H.-G. Beyer and S. Meyer-Nieberg, "Self-adaptation of evolution strategies under noisy fitness evaluations," *Genet. Programm. Evolvable Mach.*, vol. 7, pp. 295–328, 2006.

[43] J. Kybic and M. Unser, "Fast parametric elastic image registration," *IEEE Trans. Image Process.*, vol. 12, no. 11, pp. 1427–1442, Nov. 2003.

[44] S. Marsland and C. Twining, "Constructing diffeomorphic representations for the groupwise analysis of non-rigid registrations of medical images," *IEEE Trans. Med. Imag.*, vol. 23, no. 8, pp. 1006–1020, Aug. 2004.

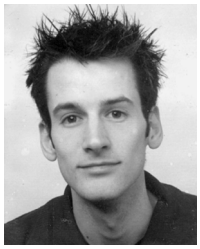
[45] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[46] 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*, vol. 21, no. 4, pp. 1428–1442, 2004.

[47] S. Hu, E. A. Hoffman, and J. M. Reinhardt, "Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images," *IEEE Trans. Med. Imag.*, vol. 20, no. 6, pp. 490–498, Jun. 2001.

[48] M. Unser, A. Aldroubi, and M. Eden, "B-Spline signal processing: Part I—Theory," *IEEE Trans. Signal Process.*, vol. 41, no. 2, pp. 821–833, Feb. 1993.

[49] 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," *Phys. Med. Biol.*, vol. 50, pp. 5869–5892, 2005.



Stefan Klein studied mechanical engineering at the University of Twente, Enschede, The Netherlands. He received the M.Sc. degree on the segmentation of fingerprint images using hidden Markov models. He is currently pursuing the Ph.D. degree in the Image Registration Group, Image Sciences Institute (ISI), Utrecht, The Netherlands.

His research focuses on optimization methods in nonrigid image registration, atlas-based segmentation, and registration problems in radiotherapy. He is one of the authors of *Elastix*, a package for medical image registration (http://www.isi.uu.nl/Elastix).

Marius Staring received the degree in applied mathematics from the University of Twente, Enschede, The Netherlands, with a thesis titled "Analysis of Quantization based Watermarking" in December 2002, a project which was carried out at Philips Research Labs (Nat. Lab.), Eindhoven, The Netherlands. He is currently pursuing the Ph.D. degree at the Image Sciences Institute (ISI), Utrecht, The Netherlands.

Since April 2003, he has been with the ISI, working on the subject of nonrigid registration of clinical images. His research focuses on registration problems that use partly rigid and partly nonrigid transformations. He is one of the authors of *Elastix*, a package for medical image registration (http://www.isi.uu.nl/Elastix).

Josien P. W. Pluim received the M.S. degree in computer science from the University of Groningen, Groningen, The Netherlands, in 1996, and the Ph.D. degree from Utrecht University, Utrecht, The Netherlands, in 2001. Her Ph.D. research was performed at the Image Sciences Institute (ISI), University Medical Center Utrecht, on the topic of mutual information based image registration.

She is now an Associate Professor at the ISI, where she heads the Image Registration Research Group.

Dr. Pluim has served as a Guest Editor for the IEEE TRANSACTIONS ON MEDICAL IMAGING and for *Medical Image Analysis*, and she is an Associate Editor for the IEEE TRANSACTIONS ON MEDICAL IMAGING. She currently chairs the SPIE Medical Imaging Image Processing Conference.