Comparison of gradient approximation techniques for optimisation of mutual information in nonrigid registration

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

Image Sciences Institute, University Medical Center Utrecht,
Q0S.459, P.O. Box 85500, 3508 GA Utrecht, The Netherlands

ABSTRACT

Nonrigid registration of medical images by maximisation of their mutual information, in combination with a deformation field parameterised by cubic B-splines, has been shown to be robust and accurate in many applications. However, the high computation time is a big disadvantage. This work focusses on the optimisation procedure. Many implementations follow a gradient-descent like approach. The time needed for computing the derivative of the mutual information with respect to the B-spline parameters is the bottleneck in this process. We investigate the influence of several gradient approximation techniques on the number of iterations needed and the computation time per iteration. Three methods are studied: a simple finite difference strategy, the so-called simultaneous perturbation method, and a more analytic computation of the gradient based on a continuous, and differentiable representation of the joint histogram. In addition, the effect of decreasing the number of image samples, used for computing the gradient in each iteration, is investigated. Two types of experiments are performed. Firstly, the registration of an image to itself, after application of a known, randomly generated deformation, is considered. Secondly, experiments are performed with 3D ultrasound brain scans, and 3D CT follow-up scans of the chest. The experiments show that the method using an analytic gradient computation outperforms the other two. Furthermore, the computation time per iteration can be extremely decreased, without affecting the rate of convergence and final accuracy, by using very few samples of the image (randomly chosen every iteration) to compute the derivative. With this approach, large data sets (256^3) can be registered within 5 minutes on a moderate PC.

Keywords: nonrigid registration, mutual information, optimisation, computation time, stochastic gradient descent

1. INTRODUCTION

This work focusses on nonrigid registration of medical images by maximisation of their mutual information, using a B-spline parameterisation of the deformation field. It has been demonstrated that this is a feasible approach.[1,2] However, the large computation time is a big disadvantage of this method. Comprehensive studies, such as lung cancer screenings, with many high resolution 3D images, ask for faster registration algorithms. Some applications, such as brain shift estimation based on intraoperatively acquired ultrasound (US) data of brain tumours,[3,4] require almost real-time registration.

Registration is usually stated as a minimisation problem:

$$\hat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu}} \left(-\mathrm{MI}(\boldsymbol{\mu})\right), \tag{1}$$

where $\mathrm{MI}(\boldsymbol{\mu})$ denotes the mutual information similarity measure as a function of the parameters, the B-spline coefficients $\boldsymbol{\mu}$. The usually high dimension of the parameter vector $\boldsymbol{\mu}$ makes this a difficult optimisation problem. To find the optimal set of parameters $\hat{\boldsymbol{\mu}}$, many existing implementations use a variant of the following iterative strategy:

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

with $\boldsymbol{d}_k$ the search direction at iteration k, and a_k a scalar gain. Common choices for the search direction are the steepest descent, the quasi-Newton, or a conjugate gradient direction. These methods have in common that

*stefan@isi.uu.nl; phone +31 30 2503186

all need (an estimate of) the derivative of the mutual information with respect to the parameters $\boldsymbol{\mu}$. The gain a_k can simply be set to a constant, defined by a decaying function of k, or determined by a 1D line search, which tries to minimise $-\mathrm{MI}(\boldsymbol{\mu}_{k+1})$ along the direction $\boldsymbol{d}_k$. As opposed to the field of *rigid* registration,[5] little research has been done on the optimisation procedure in *nonrigid* image registration problems.

In this paper the *search direction* is the object of study. We limit our attention to the steepest descent method:

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - a_k \boldsymbol{g}(\boldsymbol{\mu}_k), \tag{3}$$

where $\boldsymbol{g}(\boldsymbol{\mu}_k)$ is the derivative of the mutual information with respect to the parameters $\boldsymbol{\mu}$, at iteration k. Instead of the exact derivative $\boldsymbol{g}$, usually an approximation $\boldsymbol{g}_k$ is used. It has been proven[6,7] that, within certain constraints, this still leads to the optimum. It is important though, that the approximation is not biased; the error must be random with expectation equal to zero. In this context, the term "stochastic gradient descent" is often used. Using an approximation instead of the exact derivative can decrease the computation time per iteration, but may worsen the speed of convergence.

Three techniques for computing the gradient $\boldsymbol{g}_k$ are compared: the well-known Finite Difference (FD) technique,[7] the so-called Simultaneous Perturbation (SP) method,[8] and the approach described by Mattes et al.[1] (MA), which is an extension of the work by Thévenaz and Unser.[9] The first two methods, FD and SP, rely only on direct measurements of the mutual information. Spall[8] shows that SP in theory is superior to FD in many optimisation problems. In the last approach (MA) an analytic expression for the gradient is derived, explicitly taking into account the finite support of the B-spline basis functions that model the deformation field.

Furthermore, 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, obviously, will deteriorate the accuracy of the gradient approximation. Using fewer samples to compute the derivative of the mutual information has been proposed before,[1] but the influence on the speed of convergence in nonrigid registration has not been investigated systematically. Note, that rather than subsampling the images on a fixed grid, we mean selecting new samples randomly *every iteration*. This ensures that no bias is introduced.

In the next section additional information about the registration framework is given, the gradient estimation methods are explained in more detail, and the evaluation procedures are described. Section 3 describes the results of the experiments, and conclusions are given in Section 4.

2. METHODS

This section starts with a description of the nonrigid registration framework in which the tests are performed. In Section 2.2 the three gradient estimation methods are explained in detail. Section 2.3 focusses on the possibility of using a randomly selected set of voxels for the gradient approximations. The evaluation procedures are described in Section 2.4.

2.1. Nonrigid registration framework

In this section the various components of the nonrigid registration framework are described. The design of our algorithm is largely based on the papers of Rueckert et al.[2] and Mattes et al.[1]

As mentioned in the introduction, the registration method we focus on uses cubic B-splines to parameterise the deformation field, and *mutual information* as a similarity measure. Several implementations for the computation of mutual information can be found in the literature.[9–12] The approach described by Thévenaz and Unser[9] is used here. The joint histogram is constructed using B-spline Parzen windows, which makes it possible to formulate the mutual information as a continuous, differentiable function of the parameters describing the deformation field. In all experiments described in this paper, the joint histogram size was set to 32×32.

In order to find the correct solution to the minimisation problem (1), multiresolution strategies are often necessary. For an extensive overview on this subject we refer to Lester and Arridge.[13] Regarding the image data, we have chosen to utilise a Gaussian image pyramid in this study. The deformation model complexity is forced to 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 depend on the specific problem.

In each resolution a minimisation is performed, using the gradient descent method, as defined in Equation (3). The selection of an appropriate termination criterion for this iterative process is not straightforward, and depends on the application. In this study the minimisation is stopped after a fixed number of iterations. The gain factors a_k are defined by a function of k, specified in Section 3.1.

2.2. Gradient estimation methods

Three methods are investigated to compute the derivative $\boldsymbol{g}_k$ of the mutual information: the finite difference method (FD), a technique called simultaneous perturbation (SP), and the method described by Mattes et al.,[1] henceforth referred to as "MA".

- FD: This method, originally proposed by Kiefer and Wolfowitz,[7] is used in many applications. It is easy to implement, because only direct function measurements are used. For each element of the gradient vector, two measurements of the mutual information are needed:

$$\frac{\partial \mathrm{MI}}{\partial \left[\boldsymbol{\mu}\right]_i}(\boldsymbol{\mu}_k) \approx \left[\boldsymbol{g}_k^{\mathrm{FD}}\right]_i = \frac{\mathrm{MI}\left(\boldsymbol{\mu}_k + \Delta \boldsymbol{e}_i\right) - \mathrm{MI}\left(\boldsymbol{\mu}_k - \Delta \boldsymbol{e}_i\right))}{2\Delta} \qquad \forall i, \tag{4}$$

  where $\left[\cdot\right]_i$ represents the ith element of the vector between the brackets, Δ a small scalar, and $\boldsymbol{e}_i$ the unit vector consisting of only zeros, except for the ith element, which equals one. From this expression it is clear, that many evaluations of the mutual information are needed, when a high-dimensional parameter vector is used, i.e., when the deformation field is parameterised by a dense B-spline control point grid.

- SP: The simultaneous perturbation method, first described by Spall,[8] has the advantage that only two evaluations of the mutual information are needed to construct the complete gradient vector $\boldsymbol{g}_k$:

$$\frac{\partial \mathrm{MI}}{\partial \left[\boldsymbol{\mu}\right]_i}(\boldsymbol{\mu}_k) \approx \left[\boldsymbol{g}_k^{\mathrm{SP}}\right]_i = \frac{\mathrm{MI}\left(\boldsymbol{\mu}_k + c_k \boldsymbol{\Delta}_k\right) - \mathrm{MI}\left(\boldsymbol{\mu}_k - c_k \boldsymbol{\Delta}_k\right)}{2c_k \left[\boldsymbol{\Delta}_k\right]_i} \qquad \forall i, \tag{5}$$

  with $\boldsymbol{\Delta}_k$ the "random perturbation vector" of which each element is randomly assigned ± 1 in each iteration. The scalar c_k is defined as a function of k and the constant scalars c and γ:

$$c_k = c/(k+1)^{\gamma}. \tag{6}$$

  This is a very coarse approximation to the true derivative $\boldsymbol{g}$, but Spall[8] presents a proof of convergence pending certain assumptions. One of the assumptions is that the gain a_k in the kth iteration is computed as $a_k = a/(k+1)^{\alpha}$. In a later paper[14] Spall advises to use the following expression:

$$a_k = \frac{a}{(A+k+1)^{\alpha}}, \tag{7}$$

  with a, A, and α constants. In the same paper he suggests the settings $\alpha = 0.602$ and $\gamma = 0.101$.

- MA: This is an analytic approach that follows naturally from the definition of the mutual information as a continuous and differentiable function. In the paper by Thévenaz and Unser[9] an analytic expression for $\boldsymbol{g}_k$ is derived, valid for any differentiable parameterisation of the deformation field. Mattes et al.[1] have implemented this for the case of a deformation model based on cubic B-splines.

According to the theoretic analysis of Spall,[8] the convergence properties of FD and SP are identical, if the function to be optimised satisfies certain constraints. In combination with the much shorter computation time per iteration, this could make it a very competitive approach. However, the constraints imposed on the objective function may not be satisfied in practice. In this work it is tested if SP is applicable to the problem of B-spline based registration with mutual information as a similarity measure.

2.3. Approximation using fewer samples

Besides comparing the three methods explained in the previous section, special attention is paid to the effect of using only a small, randomly selected set of image samples for computation of the gradient.

The computation time for a single gradient evaluation is related to the amount of image samples N used and the number of B-spline coefficients M:

$$t_{\mathrm{FD}} \sim 2M(pN+q), \tag{8}$$
$$t_{\mathrm{SP}} \sim 2(pN+q), \tag{9}$$
$$t_{\mathrm{MA}} \sim rN+sM, \tag{10}$$

where p, q, r, s are constants. From these estimates it is immediately clear that, instead of using all voxels, using only a small, randomly selected subset of image samples every iteration will decrease the computation time. Of course it will also deteriorate the accuracy of the gradient approximation, which may cause a loss of convergence speed, or, even worse, diverge from the correct solution. The optimal number of samples may, amongst others, depend on the size of the image and the number of parameters of the deformation field model

Note that is very important to select *new samples in each iteration*, to prevent a systematic error in the gradient approximation.

2.4. Evaluation procedures

To compare the effects of the gradient approximation methods that were described in the previous sections, two evaluation procedures are used.

In the first evaluation procedure an image I is registered to a deformed version of itself. The applied deformation field, $\tilde{\boldsymbol{v}}$, is preferably composed of different basis function than the B-splines used in the registration. To avoid interpolation errors, the deformed version of I is not actually generated. Instead, we formulate the registration problem as:

$$\hat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu}} \left(-\mathrm{MI}(I, I'_{\boldsymbol{\mu}})\right), \tag{11}$$

where $I'_{\boldsymbol{\mu}}$ is the deformed image given the deformation parameters $\boldsymbol{\mu}$:

$$I'_{\boldsymbol{\mu}}(\boldsymbol{x}) = I\left(\boldsymbol{x} + \tilde{\boldsymbol{v}}(\boldsymbol{x}) + \boldsymbol{v}(\boldsymbol{x}, \boldsymbol{\mu})\right). \tag{12}$$

From this expression it is clear that the total deformation is composed of two parts: the known, initial deformation field $\tilde{\boldsymbol{v}}(\boldsymbol{x})$, and the deformation field parameterised by the B-spline coefficients, $\boldsymbol{v}(\boldsymbol{x}, \boldsymbol{\mu})$. Since the image I is registered to itself, the desired solution is a deformation field that is zero everywhere:

$$\tilde{\boldsymbol{v}}(\boldsymbol{x}) + \boldsymbol{v}(\boldsymbol{x}, \boldsymbol{\mu}) = 0 \qquad \forall \boldsymbol{x}. \tag{13}$$

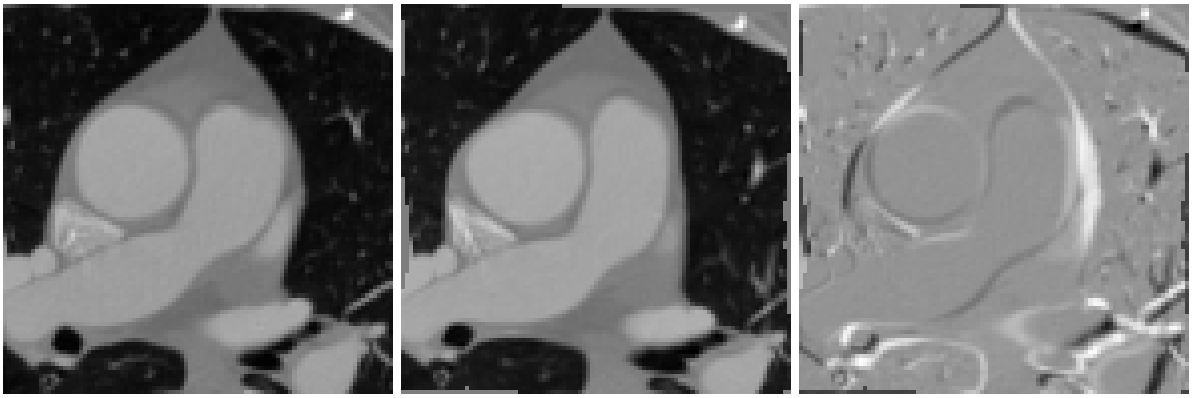
The ground truth is known, so an error measure, the average displacement error e, can be defined:

$$e(\boldsymbol{\mu}) = \frac{\sum_{\boldsymbol{x}_i \epsilon I} \|\tilde{\boldsymbol{v}}(\boldsymbol{x}_i) + \boldsymbol{v}(\boldsymbol{x}_i, \boldsymbol{\mu})\|}{|I|}, \tag{14}$$

where $\boldsymbol{x}_i$ is the position of voxel i in the image volume I, and $|I|$ the total number of voxels in I. The convergence rate is visualised by plotting this error measure as a function of the number of iterations k.

The second evaluation procedure is meant for situations where the ground truth is not known, as is common in clinical applications. One way to assess the registration results in this case is to segment corresponding structures in the images to be registered. The overlap of these corresponding structures can be computed before and after registration. An increase in overlap indicates that the alignment of the image has improved. The overlap of two segmented objects, V_1 and V_2, is defined as:

$$\mathrm{Overlap} = \frac{2 \cdot |V_1 \cap V_2|}{|V_1| + |V_2|}. \tag{15}$$



(a) (b) (c)

Figure 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 image, visualising the misalignment between (a) and (b).

3. RESULTS

Using the two evaluation procedures described in Section 2.4, three types of experiments have been performed. Firstly, the registration of a 3D CT image of the heart to a deformed version of itself is considered. Secondly, the results of registering 3D CT follow-up scans of the chest are assessed by comparing the overlap measures computed on the segmented lungs. Finally, the registration results of pre- and intra-operative 3D ultrasound brain scans are compared by computing overlap measures for segmented tumours.

All registrations have been performed using `elastix`,[15] a computer program developed by the authors. This package is largely based on the Insight Segmentation and Registration Toolkit (ITK).[16]

3.1. Registration with known ground truth

The registration experiments with known ground truth were performed on four 3D CT images of the heart. The images originated from chest scans, having an in-plane resolution of 512×512, that were cropped to cubic volumes of $194 \times 194 \times 194$ voxels containing the heart. To limit the total computation time, the resulting images were downsampled to a size of $97 \times 97 \times 97$ voxels. Voxel sizes are about 1.4 mm in all directions. Figure 1(a) shows an example of a slice of such an image.

For each image a deformation field was generated, *not* based on the previously mentioned B-spline model that is used during the registration, but composed of randomly placed Gaussian blobs. Figure 1(b) shows the same slice as in Figure 1(a), after application of the initial deformation to the volume. The difference of the image and its deformed version is displayed in Figure 1(c).

Registrations were performed using the three gradient approximation methods described in Section 2.2. Each method was tested with a varying amount of image samples used for computing the gradient. During registration the deformation field was parameterised by a $10 \times 10 \times 10$ grid of B-spline control points, which means that 3000 parameters were to be optimised. No multiresolution schemes were used, which makes comparison of the results more straightforward, and the maximum number of iterations was limited to 2048. The gain sequence of Equation (7) was adopted for all experiments. The following parameters were used for the tests with MA and FD: $a = 3200$, $A = 50$, and $\alpha = 0.602$. For SP slightly different parameters had to be used, since the method appeared to be quite sensitive to the choice of the gain sequence. The following values gave the best results: $a = 800$, $A = 200$, and $\alpha = 0.602$. Two more parameters need to be specified for SP, see Equation (6): $c = 1.4$ (equal to the voxel size) and $\gamma = 0.101$.

Figures 2-5 show the results for the four images. The error measure $e(\boldsymbol{\mu}_k)$, defined in Section 2.4, is plotted as a function of the number of iterations k. The numbers next to the labels in the graph represent the numbers

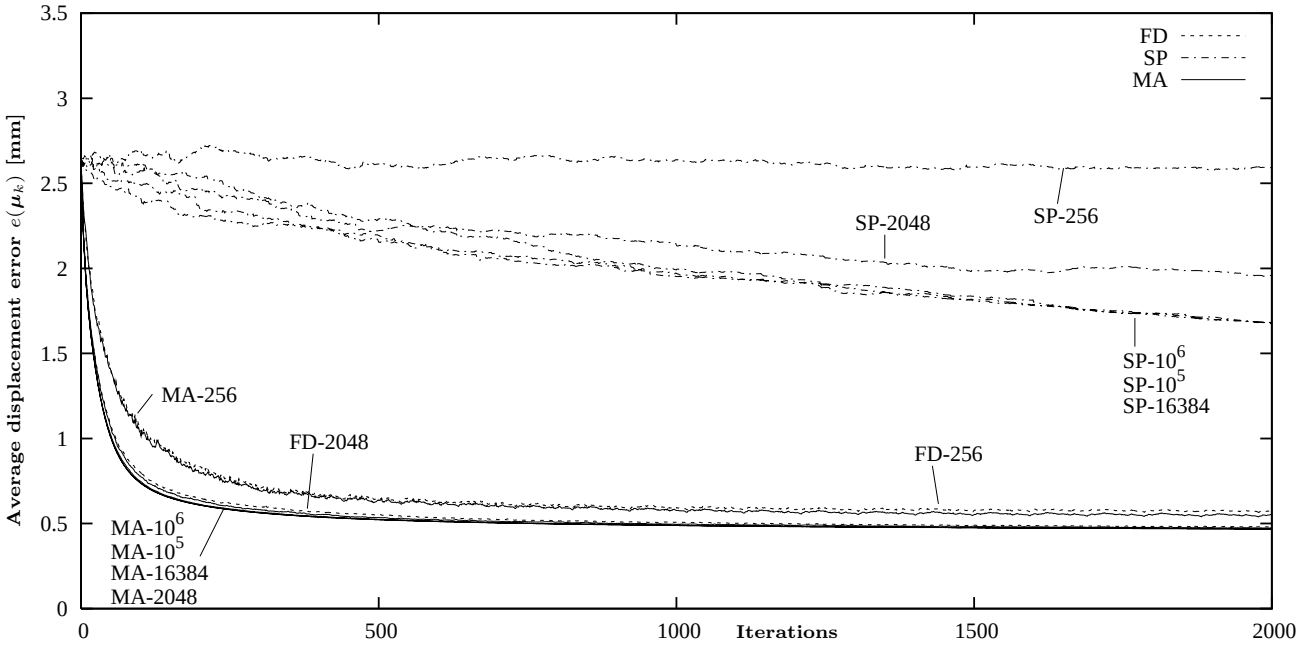


Figure 2. Results for the first heart image. The graph shows the average displacement error as a function of the number of iterations for three gradient approximation techniques: FD, SP and MA. The numbers next to the labels represent the numbers of image samples used for the gradient computations.

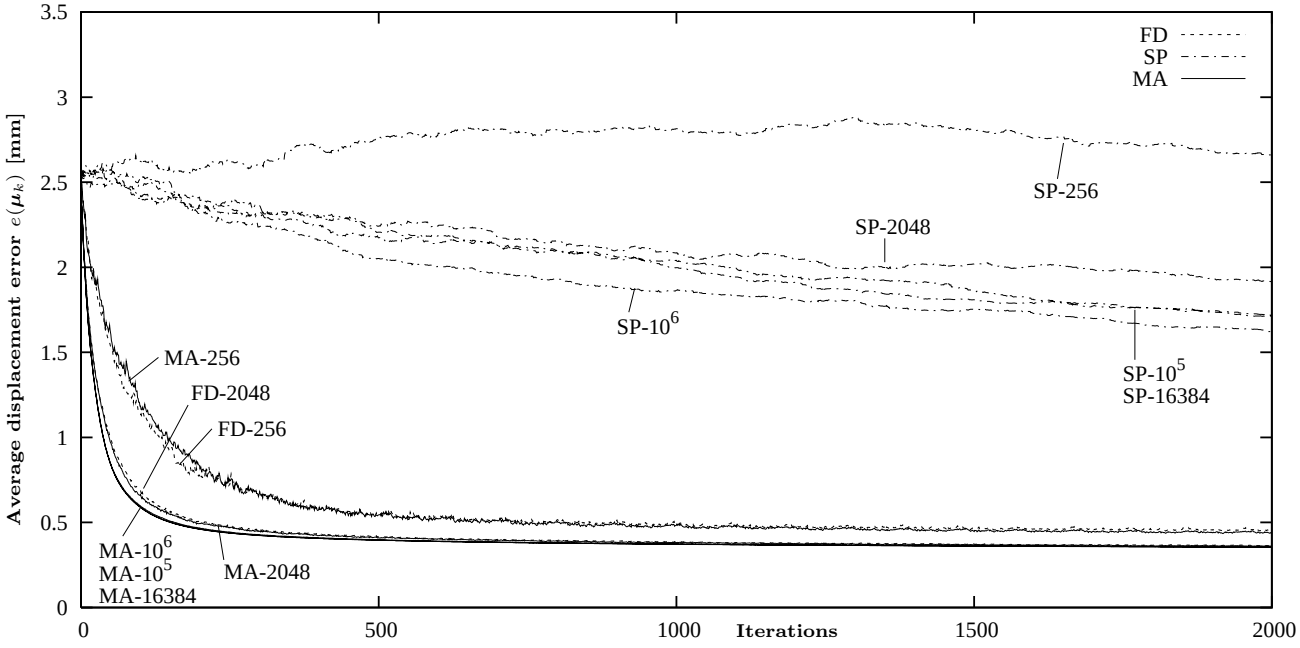


Figure 3. Results for the second heart image. The graph shows the average displacement error as a function of the number of iterations for three gradient approximation techniques: FD, SP and MA. The numbers next to the labels represent the numbers of image samples used for the gradient computations.

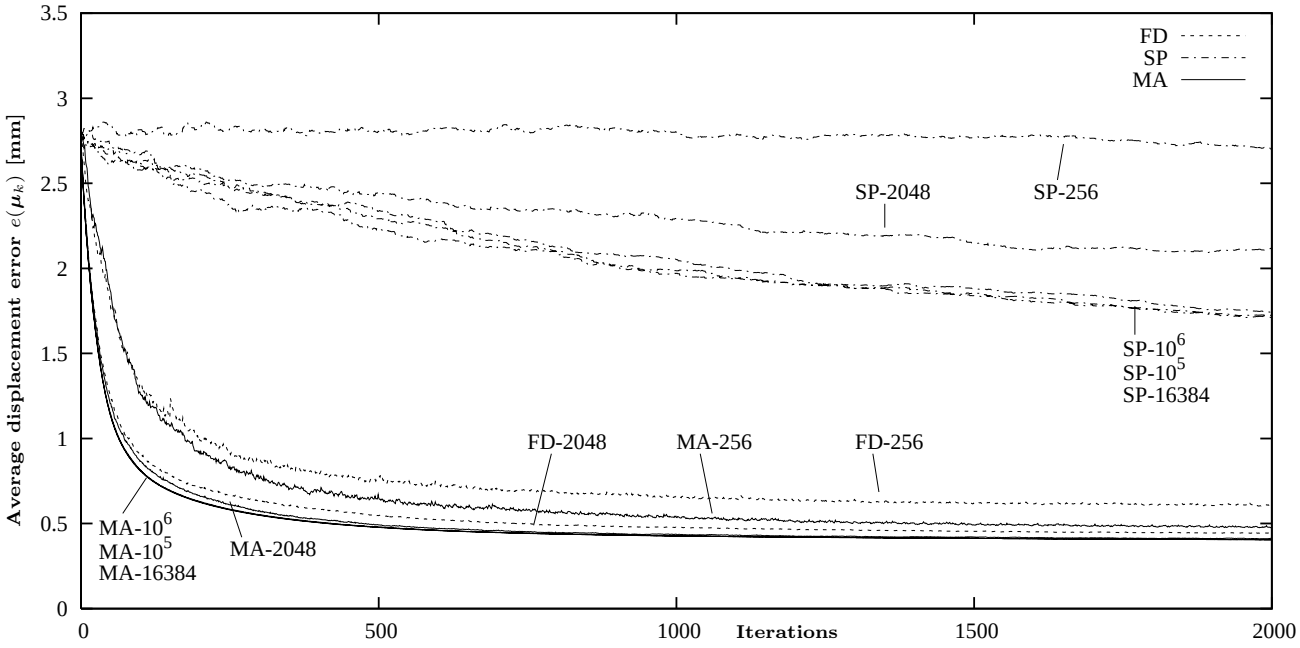


Figure 4. Results for the third heart image. The graph shows the average displacement error as a function of the number of iterations for three gradient approximation techniques: FD, SP and MA. The numbers next to the labels represent the numbers of image samples used for the gradient computations.

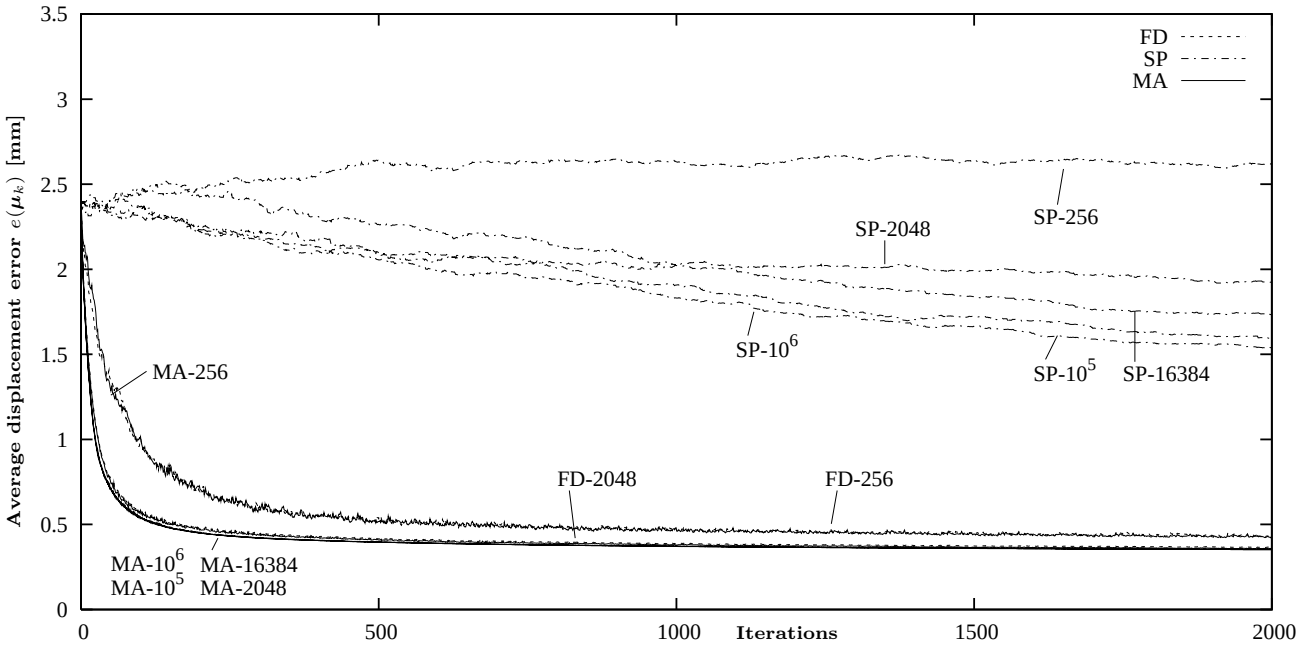


Figure 5. Results for the fourth heart image. The graph shows the average displacement error as a function of the number of iterations for three gradient approximation techniques: FD, SP and MA. The numbers next to the labels represent the numbers of image samples used for the gradient computations.

of image samples used for computation of the gradients. The FD method was only tested for 256 and 2048 image samples. Using more samples yielded unacceptable computation times on our Pentium IV 2.8 MHz.

When comparing Figures 2-5 it can be concluded that the results for the four images are very similar. Obviously, the SP approximation worsens the rate of convergence significantly. The ($\approx 4\times$) shorter computation time per iteration compared to MA does not compensate for the lower speed of convergence. The methods FD and MA have an almost equal convergence rate. However, the FD approach is much slower per iteration ($\approx 700\times$).

A remarkable result is that, for MA, the convergence properties of using all (10^6) voxels are retained when going down to only 2048 samples, which is 0.2% of the total image volume. This, in combination with the decrease in computation time per iteration gives a speed-up of a factor of approximately 500.

3.2. Overlap of segmented lungs

Three-dimensional CT is a commonly used modality for the diagnosis of lung diseases. To study the evolution of a disease in a patient it is helpful to automatically register follow-up scans. In this section a number of experiments with 3D CT chest scans are described. Assessment of registration quality is performed by segmenting the lungs, and comparing overlap measures computed on the segmented lungs. This procedure has been outlined in Section 2.4.

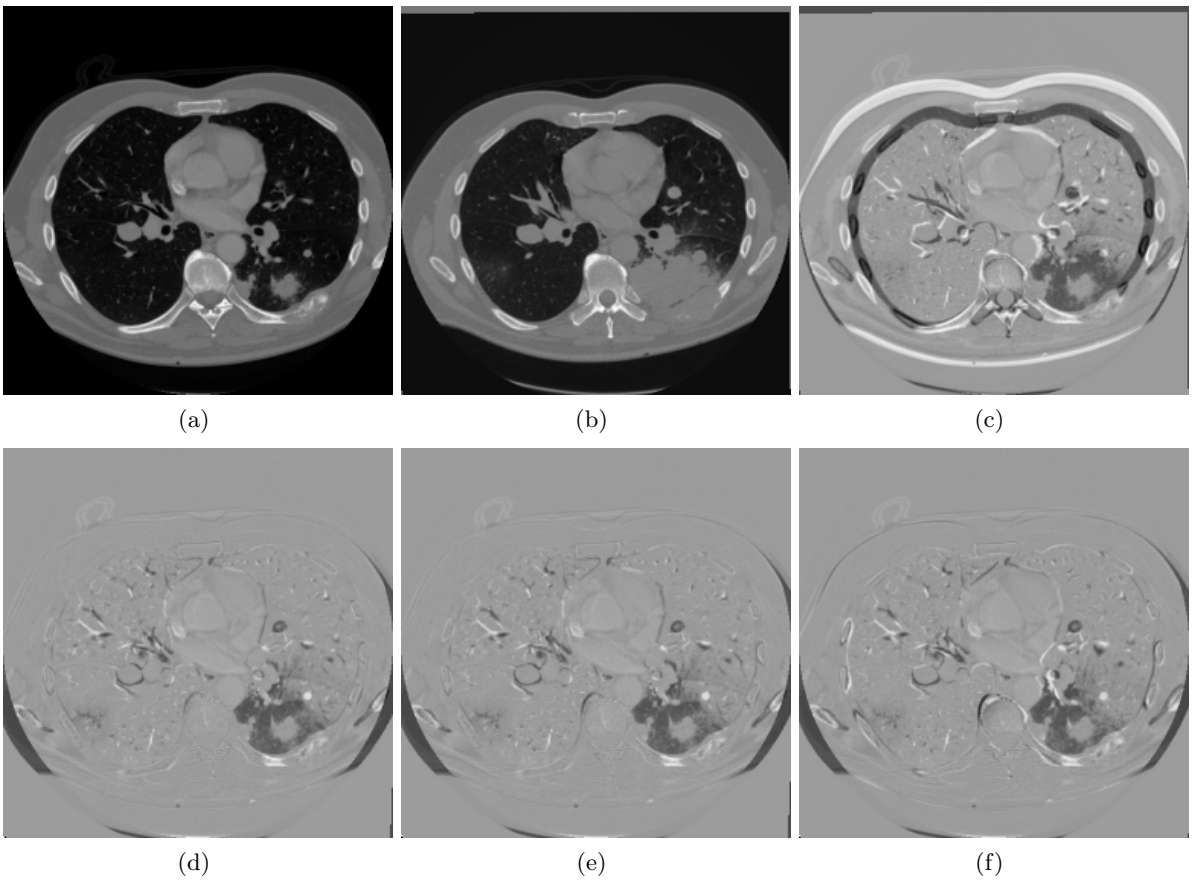


Figure 6. Examples of the CT chest data: (a) an example slice, (b) the corresponding slice in the follow-up scan after rigid registration, (c) the difference image, visualising the misalignment between (a) and (b), (d) the difference image after nonrigid registration using all voxels in each iteration, (e) the difference image after nonrigid registration using 2048 randomly selected image samples in each iteration, (f) the difference image after nonrigid registration using 256 randomly selected image samples in each iteration.

	P1	P2	P3	P4	P5
rigid	0.942	0.909	0.821	0.792	0.738
all voxels	0.976	0.973	0.943	0.958	0.955
131072	0.976	0.973	0.943	0.958	0.955
16384	0.976	0.973	0.943	0.958	0.955
2048	0.975	0.972	0.942	0.958	0.954
256	0.970	0.959	0.937	0.949	0.946

Table 1. The results of CT chest scan registration for patients P1-P5. The overlap of segmented lungs is shown after rigid registration (first row), and after nonrigid registration with a varying number of image samples used for computation of the gradient approximation. Each column contains the results for a single patient.

The images were acquired with a Philips Mx8000IDT 16-slice CT scanner. The original images, with an in-plane resolution of 512×512 and a number of slices ranging from 400 to 800, were downsampled by a factor of 2 in order to be able to register the images on a standard PC with one gigabyte of memory. The resulting voxel size is 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. In Figure 6(a) a slice is shown for illustration, next to the corresponding slice of the follow-up scan after *rigid registration* (Figure 6(b)). The difference image is displayed in Figure 6(c) to give an indication of the residual misalignment after rigid registration.

To compensate for this misalignment a *nonrigid* registration is needed. A four-level multiresolution approach was applied (see Section 2.1 for details on the utilised multiresolution scheme). At each resolution the number of iterations was fixed to 256. At the highest resolution the B-spline control point spacing was set to 22 mm, yielding a grid of about 19^3 control points; approximately 20000 parameters to optimise. The exact size of the grid varied between patients, due to the varying amounts of slices in the images. The results from the previous section show that the gradient estimation method MA is superior to FD and SP. Therefore, this method was used in the experiments, again with a varying number of image samples.

After registration the overlap of the segmented lungs can be calculated. The segmentations were made by means of a method based on the work of Hu et al.[17] In the segmentations large pulmonary vessels are not considered part of the lungs. Table 1 shows the overlap measure after rigid registration and nonrigid registration using all ($\approx 10^7$) voxels each iteration, 131072 randomly sampled voxels, 16384, 2048, and 256 samples. Each column displays the results for a single patient. The overlap measures confirm the results found in Section 3.1. The final accuracy of the nonrigid registration is not affected by the random subsampling strategy. Only when 256 samples are used, the overlap measure decreases. Visual inspection of the results supports this conclusion. In Figures 6(d), 6(e), and 6(f) an example is given of the residual misalignments after nonrigid registration using respectively all voxels, 2048 samples, and 256 samples. The difference images in Figure 6(d) and Figure 6(e) are very similar. In Figure 6(f) a slightly larger misalignment can be observed. All three images show a considerable improvement on the rigid registration (Figure 6(c)).

Surprisingly, the same minimum of 2048 samples is found as in the previous section, whilst the images considered here are almost three times larger, and the number of parameters to be optimised seven times higher.

3.3. Overlap of brain tumours

During neurosurgical interventions, 3D ultrasound (US) is an increasingly popular modality to visualise intra-operatively occurring brain deformations. The next step would be to use the ultrasound data not only for visualisation purposes, but also for automatically updating preoperatively acquired image data to the actual, intraoperative situation.[3,4] Needless to say, computation time is a critical issue for such applications. In this section the nonrigid registration of 3D US volumes is the subject of our attention. For evaluation, a similar procedure is followed as in the previous section.

The data used in this section was obtained during brain surgery, before and immediately after opening the dura. Around 100 2D scans were acquired by a free-hand sweep. For the 3D reconstruction the software package StackSX[18] was used, which combines the 2D scans using their relative positions as recorded by the

(a) (b) (c)

Figure 7. Examples of the US brain data: (a) a slice from a patient, immediately after opening of the dura, (b) the corresponding slice in the image created before opening the dura (after rigid registration), (c) the difference image, visualising the misalignment between (a) and (b) around the tumour.

	P1	P2	P3
rigid	0.809	0.900	0.851
all voxels	0.822	0.932	0.888
131072	0.822	0.932	0.888
16384	0.821	0.932	0.888
2048	0.817	0.932	0.887
256	0.797	0.930	0.884

Table 2. The results of US brain scan registration for patients P1-P3. The overlap of segmented tumours is shown after rigid registration (first row), and after nonrigid registration with a varying number of image samples used for computation of the gradient approximation. Each column contains the results for a single patient.

neuro-navigation system. The resulting 3D images have a size of $300 \times 300 \times 200$ voxels, with isotropic voxels of $0.5 \times 0.5 \times 0.5$ mm. Figure 7(a) shows an example slice from an image acquired after opening the dura. The tumour is clearly visible. In Figure 7(b) the corresponding slice in the image obtained before opening the dura is shown. Only a rigid registration has been performed between these images. In the difference image, displayed in Figure 7(c), it can be seen that the tumour alignment is already quite good. However, the rigid transformation model could not completely recover the brain shift. A nonrigid registration is necessary.

For the nonrigid registration a three-level multiresolution scheme was used. At the highest resolution, the deformation field was modelled by a $41 \times 41 \times 28$ grid of control points, having a spacing of 4 mm. The optimisation at each resolution was stopped after 256 iterations. To estimate the derivative of the mutual information with respect to the B-spline coefficients, the MA approach was used. Only voxels inside the US beam were taken into account in the registration.

For three patients the 3D US scans were registered using different amounts of samples. In all images the tumours were manually segmented by an experienced person, which allows the computation of the overlap measure, defined in Section 2.4. The results are presented in Table 2. The same pattern can be observed as in the former two sections. Even when only 2048 samples are used in each iteration, the accuracy of the registration is retained. The results are less consistent for patient 1. This may be a result of the quality of the ultrasound scans, which makes it difficult to segment the tumours.

4. CONCLUSION

We investigated the influence of several gradient approximation techniques on the convergence speed and the computation time per iteration. The experiments performed clearly show that extremely few samples are needed for obtaining an acceptable approximation of the gradient. The minimum number was observed to lie around 2048. Even in the case of large 3D images, in combination with many parameters to be optimised, the speed of convergence and the final accuracy did not suffer at all from this approach. Accelerations up to a factor of 500 could easily be achieved with this strategy.

A big difference between the three gradient computation methods was observed. The finite difference method converges fast, but the large number of parameters to be optimised causes a very high computation time per iteration. The computation time of gradient approximation by simultaneous perturbation of the parameters does not depend on the number of B-spline coefficients, and is therefore faster per iteration, but its convergence properties are much worse. The expression for the derivative described by Mattes et al. makes efficient use of the small support of the B-spline basis functions modelling the deformation field. This, in combination with the fast convergence, makes it the preferred method. In applications where the deformation field is modelled by basis functions with a global support, such as thin-plate splines, the simultaneous perturbation method may become more advantageous.

Future extensions of this work include the investigation of different methods to set the gain a_k. Instead of a decaying function of k, other functions can be used, which may improve the rate of convergence. In the literature on stochastic gradient descent optimisation a so-called "search-then-converge" strategy[19] has been suggested for example. Another approach would be to use a one-dimensional line search, which optimises the gain every iteration.

A second important topic is the choice of search direction. In this paper we focussed on gradient descent methods. As mentioned in Section 1 other options include the quasi-Newton search direction, and the conjugate gradient. We aim to evaluate if these approaches can benefit from the random subsampling technique as well.

ACKNOWLEDGMENTS

Funding of this research has been provided by the Netherlands Organisation for Scientific Research (NWO). 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 `www.itk.org`.

REFERENCES

1. 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), pp. 120–128, 2003.
2. 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), pp. 712–721, 1999.
3. X. Pennec, P. Cachier, and N. Ayache, "Tracking brain deformations in time sequences of 3D US images," *Pattern Recognition Letters* **24**(4-5), pp. 801–813, 2003.
4. M. M. J. Letteboer, P. Willems, M. A. Viergever, and W. J. Niessen, "Non-rigid registration of 3D Ultrasound Images of Brain Tumours Acquired during Neurosurgery," in *Medical Image Computing and Computer-Assisted Intervention*, R. E. Ellis and T. M. Peters, eds., *Lecture Notes in Computer Science* **2879**, pp. 408–415, Springer, 2003.
5. 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), pp. 373–386, 1999.
6. H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics* **22**(3), pp. 400–407, 1951.
7. J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *The Annals of Mathematical Statistics* **23**(3), pp. 462–466, 1952.

8. J. C. Spall, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Transactions on Automatic Control* **37**(3), pp. 332–341, 1992.
9. P. Thévenaz and M. Unser, "Optimization of mutual information for multiresolution image registration," *IEEE Transactions on Image Processing* **9**(12), pp. 2083–2099, 2000.
10. 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), pp. 187–198, 1997.
11. 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), pp. 163–175, 1996.
12. P. Viola and W. M. Wells III, "Alignment by maximization of mutual information," *International Journal of Computer Vision* **24**(2), pp. 137–154, 1997.
13. H. Lester and S. R. Arridge, "A survey of hierarchical non-linear medical image registration," *Pattern Recognition* **32**(1), pp. 129–149, 1999.
14. J. C. Spall, "Implementation of the simultaneous perturbation method for stochastic optimization," *IEEE Transactions on Aerospace and Electronic Systems* **34**, pp. 817–823, 1998.
15. S. Klein and M. Staring, "`elastix`." `http://www.isi.uu.nl/Elastix`.
16. "Insight Segmentation and Registration Toolkit (ITK)." `http://www.itk.org`.
17. 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), pp. 490–498, 2001.
18. R. W. Prager, A. Gee, and L. Berman, "Stradx: real-time acquisition and visualization of freehand 3D ultrasound," *Medical Image Analysis* **3**(2), pp. 129–140, 1999.
19. C. Darken and J. Moody, "Towards faster stochastic gradient search," in *Advances in Neural Information Processing Systems*, J. E. Moody, S. J. Hanson, and R. P. Lipmann, eds., **4**, pp. 1009–1016, Morgan Kaufmann Publishers, San Mateo, CA, 1991.