Medical Image Analysis (1999) volume 3, number 4, pp 373–386

Comparative evaluation of multiresolution optimization strategies for multimodality image registration by maximization of mutual information

Frederik Maes[†]*, Dirk Vandermeulen and Paul Suetens

Katholieke Universiteit Leuven, Department of Electrical Engineering (ESAT-PSI), Kard. Mercierlaan 94, B-3001 Heverlee, Belgium

Abstract

Maximization of mutual information of voxel intensities has been demonstrated to be a very powerful criterion for three-dimensional medical image registration, allowing robust and accurate fully automated affine registration of multimodal images in a variety of applications, without the need for segmentation or other preprocessing of the images. In this paper, we investigate the performance of various optimization methods and multiresolution strategies for maximization of mutual information, aiming at increasing registration speed when matching large high-resolution images. We show that mutual information is a continuous function of the affine registration parameters when appropriate interpolation is used and we derive analytic expressions of its derivatives that allow numerically exact evaluation of its gradient. Various multiresolution gradient- and non-gradient-based optimization strategies, such as Powell, simplex, steepest-descent, conjugate-gradient, quasi-Newton and Levenberg–Marquardt methods, are evaluated for registration of computed tomography (CT) and magnetic resonance images of the brain. Speed-ups of a factor of 3 on average compared to Powell's method at full resolution are achieved with similar precision and without a loss of robustness with the simplex, conjugate-gradient and Levenberg–Marquardt method using a two-level multiresolution scheme. Large data sets such as $256^2 \times 128$ MR and $512^2 \times 48$ CT images can be registered with subvoxel precision in $<$5 min CPU time on current workstations.

Keywords: gradient, multiresolution, mutual information, optimization, registration

Received August 21, 1998; revised January 4, 1999; accepted January 29, 1999

1. INTRODUCTION

Maximization of mutual information (MMI) has recently been proposed as a new approach for multimodal medical image registration (Viola and Wells, 1995; Collignon *et al.*, 1995; Studholme *et al.*, 1996; Wells *et al.*, 1996; Maes *et al.*, 1997; Meyer *et al.*, 1997). The method applies the concept of mutual information (MI) to measure the statistical dependence between the image intensities of corresponding voxels in both images, which is assumed to be maximal if the images are geometrically aligned. Because no assumptions are made regarding the nature of this dependence and no limiting constraints are imposed on the image content of the modalities involved, MMI is a very general and powerful criterion, allowing robust, fully automated affine registration of multimodal images with different contrast and resolution in a variety of applications without the need for segmentation or other preprocessing (Maes, 1998). Comparative studies of state-of-the-art algorithms for multimodality image registration have demonstrated the superior performance of MMI for registration of computed tomography (CT), magnetic resonance (MR) and positron-emission tomography (PET) images of the brain (Studholme *et al.*, 1996; West *et al.*, 1997). Several groups have experimented with

*Corresponding author
(e-mail: Frederik.Maes@uz.kuleuven.ac.be)

[†] Frederik Maes is Postdoctoral Fellow of the Fund for Scientific Research, Flanders, Belgium.

the MI criterion and have presented different optimization strategies. Collignon *et al.* (1995) and Maes *et al.* (1997) use Powell's direction set method (Press *et al.*, 1992) to maximize MI, which is evaluated from the joint image intensity histogram constructed from the overlapping volume of both images. Wells *et al.* (1996) use a gradient-descent approach, computing a stochastic approximation for the gradient of the MI criterion from an estimate of the joint histogram derived by Parzen windowing (Duda and Hart, 1973) from a very limited number of samples. Studholme *et al.* (1996) use a multiresolution technique to speed up a heuristic search procedure in which the parameters are changed iteratively one by one by small amounts so as to maximize MI. Meyer *et al.* (1997) use a simplex search method to optimize the parameters of an affine or thin-plate spline-warped registration. None of these authors have systematically investigated the performance of the optimization method used to maximize the MI measure.

In this paper we evaluate various gradient- and non-gradient-based optimization strategies for MMI, aiming to increase the registration speed when matching large high-resolution images. We show that the MI criterion varies smoothly as a function of the 12 affine registration parameters when appropriate interpolation is used (Maes *et al.*, 1997) and its gradient can be computed exactly, using analytic expressions for the partial derivatives of MI with respect to the registration parameters. Because the complexity of evaluating MI or its gradient is proportional to the number of samples in the image, large speed-ups are possible using a multiresolution approach, starting the optimization at a low resolution using a coarsely sampled image and continuing at a higher resolution by increasing the number of samples as the optimization proceeds. Various optimization strategies, including Powell, downhill simplex, steepest gradient-descent, conjugate-gradient, quasi-Newton and Levenberg–Marquardt (Press *et al.*, 1992), have been implemented and evaluated for registration of high-resolution CT and MR images of the brain. Our results demonstrate that, compared to Powell's method applied at full resolution as in Maes *et al.* (1997), speed-ups by a factor of three on average are achieved systematically with a similar precision and without loss of robustness with the simplex, conjugate-gradient and Levenberg–Marquardt methods using a two-level multiresolution scheme. Large data sets such as $256^2 \times 128$ MR and $512^2 \times 48$ CT images can be registered with subvoxel precision in less than 5 min CPU time on a Silicon Graphics (SGI) workstation.

This paper is organized as follows. The mutual information registration criterion is presented briefly in Section 2. Analytic expressions for the gradient of MI w.r.t. the affine registration parameters are derived in Section 3. Section 4 describes the various optimization strategies for MMI that were compared in this study. Section 5 evaluates the performance of these methods for matching of CT and MR brain images. These results are discussed in Section 6. Conclusions are presented in Section 7.

2. THE MUTUAL INFORMATION CRITERION

The affine geometric transformation $\mathbf{T}_{\mathcal{FR}}(\boldsymbol{\alpha})$ of image coordinates $\mathbf{p}$ in the floating image $\mathcal{F}$ into image coordinates $\mathbf{q}$ in the reference image $\mathcal{R}$ is given by

$$\mathbf{q} = \mathbf{T}_{\mathcal{FR}}(\boldsymbol{\alpha}) \cdot \mathbf{p} = (\mathbf{T}^{-1}_{i\to w,\mathcal{R}} \cdot \mathbf{T}_w(\boldsymbol{\alpha}) \cdot \mathbf{T}_{i\to w,\mathcal{F}}) \cdot \mathbf{p}$$

where $\mathbf{T}_{i\to w,\mathcal{F}}$ and $\mathbf{T}_{i\to w,\mathcal{R}}$ are 4×4 image-to-world coordinate transformation matrices that take the voxel sizes and orientation of each of the images into account. We define the x, y and z axes of the world coordinate frame in each image relative to the patient as the right-to-left, anterior-to-posterior and vertical axis respectively. $\mathbf{T}_w(\boldsymbol{\alpha})$ is the 4×4 affine world-to-world coordinate transformation matrix determined by the 12 affine transformation parameters $\boldsymbol{\alpha} = \{\alpha_\iota\} = \{\mathbf{t}, \boldsymbol{\phi}, \mathbf{g}, \mathbf{s}\}$:

$$\mathbf{T}_w(\boldsymbol{\alpha}) = \mathbf{T}_t(\mathbf{t}) \cdot \mathbf{T}_r(\boldsymbol{\phi}) \cdot \mathbf{T}_g(\mathbf{g}) \cdot \mathbf{T}_s(\mathbf{s}) \tag{1}$$

with the 4×4 matrices $\mathbf{T}_t$, $\mathbf{T}_r$, $\mathbf{T}_g$ and $\mathbf{T}_s$ representing three-dimensional (3-D) translation, rotation, skew and scaling respectively.

The joint image intensity histogram $\mathbf{H} = \{h_{fr}\}$ of the overlapping volume of images $\mathcal{F}$ and $\mathcal{R}$ with image intensities $\{f\}$ and $\{r\}$ respectively is constructed by transforming samples $\{\mathbf{p}_k\}$ in image $\mathcal{F}$ into samples $\{\mathbf{q}_k = \mathbf{T}_{\mathcal{FR}}(\boldsymbol{\alpha}) \cdot \mathbf{p}_k\}$ in image $\mathcal{R}$, and binning all pairs of corresponding voxel intensities $\{(f_k = \mathcal{F}(\mathbf{p}_k), r_k = \mathcal{R}(\mathbf{q}_k))\}$ after interpolation in image $\mathcal{R}$. The mutual information I of $\mathcal{F}$ and $\mathcal{R}$ is computed from $\mathbf{H}$:

$$I = \frac{1}{N} \sum_{f,r} h_{fr} \log_2 \frac{N \cdot h_{fr}}{h_f \cdot h_r}$$

with $h_f = \sum_r h_{fr}$, $h_r = \sum_f h_{fr}$ and $N = \sum_{f,r} h_{fr}$. $\mathbf{H}$ and I are functions of $\mathbf{T}_{\mathcal{FR}}$ and hence of $\boldsymbol{\alpha}$. The mutual information registration criterion states that $\mathcal{F}$ and $\mathcal{R}$ are properly aligned for the registration parameters $\boldsymbol{\alpha}^*$ that maximize I:

$$\boldsymbol{\alpha}^* = \arg\max I(\boldsymbol{\alpha})$$

3. THE GRADIENT OF MUTUAL INFORMATION

If the derivatives $\partial\mathbf{H}/\partial\alpha_\iota = \{\partial h_{fr}/\partial\alpha_\iota\}$ of the joint histogram $\mathbf{H}$ w.r.t. the affine registration parameters α_ι exist, the gradient $\nabla I(\boldsymbol{\alpha})$ of mutual information I can be written as

$$\nabla I(\boldsymbol{\alpha}) = \left\{ \frac{\partial I}{\partial \alpha_\iota} \right\}$$

$$\frac{\partial I}{\partial \alpha_\iota} = \sum_{f,r} \frac{\mathrm{d}I}{\mathrm{d}h_{fr}} \cdot \frac{\partial h_{fr}}{\partial \alpha_\iota} = \sum_{f,r} \frac{1}{N} \left(\log_2 \frac{N \cdot h_{fr}}{h_f \cdot h_r} - I \right) \frac{\partial h_{fr}}{\partial \alpha_\iota}.$$

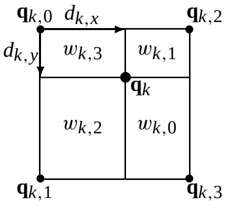
Each gradient component ι is thus expressed as the sum over all histogram entries of the change in each entry when changing α_ι, weighted by the influence of this change on I.

The dependence of $\mathbf{H}$ on $\boldsymbol{\alpha}$ depends on the interpolation scheme and the binning strategy that is used to construct $\mathbf{H}$. Various interpolation schemes can be considered (Maes *et al.*, 1997). Intensity interpolation methods, such as nearest-neighbour (NN) and trilinear (TRI) interpolation, estimate the intensity r at the transformed position $\mathbf{q}$ and for each sample in the volume of overlap of both images a single bin in their joint histogram is incremented with one. But this results in discontinuous changes in the histogram for small changes in the registration parameters, due to abrupt changes in the NN when using NN interpolation and due to the discrete nature of the binning process when using TRI or higher-order interpolation. Trilinear partial volume distribution (PV) interpolation on the other hand, as introduced in Collignon *et al.* (1995) and Maes *et al.* (1997), assures that $\mathbf{H}$ varies continuously with $\boldsymbol{\alpha}$, such that the histogram derivatives $\partial\mathbf{H}/\partial\alpha_\iota$ exist almost everywhere when PV interpolation is used (they can be infinite) and can be computed exactly using analytic expressions. We first formalize mathematically the PV interpolation scheme and then derive expressions for $\partial\mathbf{H}/\partial\alpha_\iota$ and $\partial I/\partial\alpha_\iota$.

3.1. Computing H using PV interpolation

The images $\mathcal{F}$ and $\mathcal{R}$ are first linearly rescaled such that the intensities $\{f\}$ in $\mathcal{F}$ and $\{r\}$ in $\mathcal{R}$ satisfy $1 \leq f \leq n_\mathcal{F} - 1$ and $1 \leq r \leq n_\mathcal{R} - 1$, with $n_\mathcal{F}$ and $n_\mathcal{R}$ the number of histogram entries for image $\mathcal{F}$ and $\mathcal{R}$ respectively. The $n_\mathcal{F} \times n_\mathcal{R}$ joint histogram $\mathbf{H}$ is then computed using PV interpolation by distributing the contribution of each sample taken from $\mathcal{F}$ over the up to eight different entries in $\mathbf{H}$ that correspond to the intensity of this sample in $\mathcal{F}$ and the intensities of the eight nearest neighbours of its transformation in $\mathcal{R}$, using the same weights as used for trilinear interpolation. This is depicted schematically in Figure 1.

Let k be a voxel of image $\mathcal{F}$ at position $\mathbf{p}_k$. Let $\mathbf{q}_k = \mathbf{T}_{\mathcal{F}\mathcal{R}} \cdot \mathbf{p}_k$ be the transformed position of voxel k in image $\mathcal{R}$. Let $l_{k,m}$, with $m = 4\alpha_m + 2\beta_m + \gamma_m$, $\alpha_m, \beta_m, \gamma_m \in \{0, 1\}$, $m = 0, 1, 2, \ldots, 7$, denote the eight NN voxels of $\mathbf{q}_k$ in image $\mathcal{R}$, at positions $\mathbf{q}_{k,m}$ such that $\mathbf{q}_{k,m_1} - \mathbf{q}_{k,m_2} = \mathbf{n}_{m_1} - \mathbf{n}_{m_2}$ with $\mathbf{n}_m = (\alpha_m, \beta_m, \gamma_m)$. Let $\mathbf{d}_k$ be the distance of $\mathbf{q}_k$ to $\mathbf{q}_{k,0}$: $\mathbf{d}_k = \mathbf{q}_k - \mathbf{q}_{k,0} = (d_{kx}, d_{ky}, d_{kz})$ with $0 \leq d_{ki} \leq 1$.



$$\begin{aligned}
&w_{k,0} = (1 - d_{k,x}) \cdot (1 - d_{k,y})\\
&w_{k,1} = (1 - d_{k,x}) \cdot d_{k,y}\\
&w_{k,2} = d_{k,x} \cdot (1 - d_{k,y})\\
&w_{k,3} = d_{k,x} \cdot d_{k,y}\\
&\textstyle\sum_m w_{k,m} = 1\\
&f_k = \mathcal{F}(\mathbf{p}_k), r_{k,m} = \mathcal{R}(\mathbf{q}_{k,m})\\
&\forall m\colon H(f_k, r_{k,m}) \leftarrow H(f_k, r_{k,m}) + w_{k,m}
\end{aligned}$$

Figure 1. The PV interpolation scheme in two dimensions. The contribution of each sample k at position $\mathbf{p}_k$ in image $\mathcal{F}$ and $\mathbf{q}_k$ in image $\mathcal{R}$ to the joint histogram $\mathbf{H}$ of $\mathcal{F}$ and $\mathcal{R}$ is distributed over all entries $(\mathcal{F}(\mathbf{p}_k), \mathcal{R}(\mathbf{q}_{k,m}))$ in $\mathbf{H}$, with $\mathbf{q}_{k,m}$ the NN of $\mathbf{q}_k$ in $\mathcal{R}$, using the same distribution weights $w_{k,m}$ as for trilinear interpolation, proportional to the areas indicated in the figure.

The trilinear partial volume distribution weights $w_{k,m}$ of $\mathbf{q}_k$ w.r.t. its neighbours $\mathbf{q}_{k,m}$ are then given by

$$\begin{aligned}
w_{k,0} &= (1 - d_{kx}) \cdot (1 - d_{ky}) \cdot (1 - d_{kz})\\
w_{k,1} &= (1 - d_{kx}) \cdot (1 - d_{ky}) \cdot d_{kz}\\
w_{k,2} &= (1 - d_{kx}) \cdot d_{ky} \cdot (1 - d_{kz})\\
w_{k,3} &= (1 - d_{kx}) \cdot d_{ky} \cdot d_{kz}\\
w_{k,4} &= d_{kx} \cdot (1 - d_{ky}) \cdot (1 - d_{kz})\\
w_{k,5} &= d_{kx} \cdot (1 - d_{ky}) \cdot d_{kz}\\
w_{k,6} &= d_{kx} \cdot d_{ky} \cdot (1 - d_{kz})\\
w_{k,7} &= d_{kx} \cdot d_{ky} \cdot d_{kz}.
\end{aligned} \tag{2}$$

Let f_k be the intensity of image $\mathcal{F}$ at voxel k. Let $r_{k,m}$ be the intensity of image $\mathcal{R}$ at voxel $l_{k,m}$ if $\mathbf{q}_{k,m}$ falls inside the volume of $\mathcal{R}$ and 0 otherwise. For each sample k in $\mathcal{F}$, $\mathbf{H}$ is updated at all eight entries $h_{f_k r_{k,m}}$ with the weight $w_{k,m}$:

$$h_{f_k r_{k,m}} \leftarrow h_{f_k r_{k,m}} + w_{k,m}, \qquad m = 0, 1, 2, \ldots, 7.$$

Each entry h_{fr} in the histogram $\mathbf{H}$ is thus built up as the sum over all samples k of all fractions $w_{k,m}$:

$$\mathbf{H}(f, r) = \sum_k \sum_{m=0}^{7} w_{k,m} \cdot \delta(f - f_k, r - r_{k,m}) \tag{3}$$

with $\delta(x, y)$ the discrete unit pulse.

Entries of $\mathbf{H}(f, r)$ corresponding to $f = 0$ or $r = 0$ are not taken into account when computing I to exclude samples that fall outside the volume of overlap. Samples k in image $\mathcal{F}$ for which some of its neighbours $l_{k,m}$ in image $\mathcal{R}$ fall outside the volume of $\mathcal{R}$ are thus only partially taken into account and the contribution of a sample moving outside of the volume of overlap when the registration parameters are changed decreases gracefully to zero. Because all fractions $w_{k,m}$ vary smoothly with the transformed position $\mathbf{q}_k$, which itself varies smoothly with the transformation parameters $\boldsymbol{\alpha}$, the histogram $\mathbf{H}$ and hence the mutual information I are continuous functions of $\boldsymbol{\alpha}$.

3.2. Computing derivatives of H using PV interpolation

The derivatives of $\mathbf{H}$ w.r.t. the affine transformation parameters α_ι can be expressed as a linear combination of the derivatives of $\mathbf{H}$ w.r.t. the 12 elements $T_{\mathcal{FR},ij}$, $i = 0, 1, 2$, $j = 0, 1, 2, 3$, of the image-to-image transformation matrix itself:

$$\frac{\partial \mathbf{H}}{\partial \alpha_\iota} = \sum_{ij} \frac{\partial \mathbf{T}_{\mathcal{FR},ij}}{\partial \alpha_\iota} \cdot \frac{\partial \mathbf{H}}{\partial T_{\mathcal{FR},ij}}.$$

The derivatives $\partial \mathbf{T}_{\mathcal{FR}}/\partial \alpha_\iota$ of the image-to-image transformation matrix $\mathbf{T}_{\mathcal{FR}}$ with respect to its parameters α_ι are itself 4×4 matrices given by

$$\frac{\partial \mathbf{T}_{\mathcal{FR}}}{\partial \alpha_\iota} = \mathbf{T}^{-1}_{i \to w,\mathcal{R}} \cdot \frac{\partial \mathbf{T}_w}{\partial \alpha_\iota} \cdot \mathbf{T}_{i \to w,\mathcal{F}}. \tag{4}$$

Expressions for $\partial \mathbf{T}_w/\partial \alpha_\iota$ can be computed straightforwardly for each of the translation, rotation, skew and scale parameters from expression (1) given above (Maes, 1998).

Using expression (3) for $\mathbf{H}$, the derivatives of $\mathbf{H}$ w.r.t. $T_{\mathcal{FR},ij}$ can be computed using PV interpolation from the derivatives of the weights $w_{k,m}$ w.r.t. $T_{\mathcal{FR},ij}$:

$$\frac{\partial \mathbf{H}(f,r)}{\partial T_{\mathcal{FR},ij}} = \sum_k \sum_{m=0}^{7} \frac{\partial w_{k,m}}{\partial T_{\mathcal{FR},ij}} \cdot \delta(f - f_k, r - r_{k,m}). \tag{5}$$

Thus, the derivative $\partial \mathbf{H}/\partial T_{\mathcal{FR},ij}$ is itself a histogram which can be computed using the PV interpolation scheme similarly to the way $\mathbf{H}$ itself is computed, but updating each entry with $\partial w_{k,m}/\partial T_{\mathcal{FR},ij}$ instead of with $w_{k,m}$.

Expressions for $\partial w_{k,m}/\partial T_{\mathcal{FR},ij}$ can be obtained as follows:

$$\frac{\partial w_{k,m}}{\partial T_{\mathcal{FR},ij}} = \sum_s \frac{\partial w_{k,m}}{\partial d_{ks}} \cdot \frac{\partial d_{ks}}{\partial T_{\mathcal{FR},ij}}. \tag{6}$$

The derivatives $\partial w_{k,m}/\partial d_{ks}$ of $w_{k,m}$ w.r.t. d_{kx}, d_{ky} and d_{kz} are simply computed from the expressions (2) for $w_{k,m}$ given above, while the derivatives $\partial \mathbf{d}_k/\partial T_{\mathcal{FR},ij}$ of $\mathbf{d}_k$ w.r.t. the elements $T_{\mathcal{FR},ij}$ of the image-to-image transformation are

$$\frac{\partial \mathbf{d}_k}{\partial T_{\mathcal{FR},ij}} = \frac{\partial \mathbf{q}_k}{\partial T_{\mathcal{FR},ij}} = \frac{\partial \mathbf{T}_{\mathcal{FR}}}{\partial T_{\mathcal{FR},ij}} \cdot \mathbf{p}_k.$$

Due to the sparseness of $\partial \mathbf{T}_{\mathcal{FR}}/\partial T_{\mathcal{FR},ij}$, expression (6) can be simplified:

$$\frac{\partial d_{ks}}{\partial T_{\mathcal{FR},ij}} = \begin{cases} p_{kj} & \text{if } s = i \\ 0 & \text{otherwise} \end{cases}$$

with $p_{k3} = 1$, such that

$$\frac{\partial w_{k,m}}{\partial T_{\mathcal{FR},ij}} = \frac{\partial w_{k,m}}{\partial d_{ki}} \cdot p_{kj}. \tag{7}$$

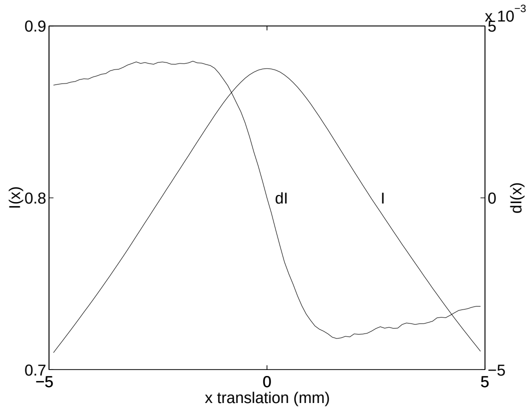


Figure 2. Traces of I and $\partial I/\partial t_x$ using PV interpolation for right-to-left translation of a CT and an MR image of the brain around the registered position.

3.3. Computing derivatives of I using PV interpolation

Combining the expressions derived above, the derivatives of $I(\boldsymbol{\alpha})$ are given by

$$\begin{aligned} \frac{\partial I}{\partial \alpha_\iota} &= \sum_{f,r} \frac{\mathrm{d}I}{\mathrm{d}h_{fr}} \cdot \frac{\partial h_{fr}}{\partial \alpha_\iota} \\ &= \sum_{f,r} \frac{1}{N} \left(\log_2 \frac{N \cdot h_{fr}}{h_f \cdot h_r} - I \right) \\ &\quad \times \left(\sum_{ij} \frac{\partial \mathbf{T}_{\mathcal{FR},ij}}{\partial \alpha_\iota} \cdot \frac{\partial \mathbf{H}(f,r)}{\partial T_{\mathcal{FR},ij}} \right). \end{aligned} \tag{8}$$

The computation of the derivatives $\partial I/\partial \alpha_\iota$ thus involves:

- computing the 12 histograms $\partial \mathbf{H}/\partial T_{\mathcal{FR},ij}$ using expressions (5) and (7); these histogram derivatives can be computed together with $\mathbf{H}$ by scanning the image $\mathcal{F}$ once;
- computing the derivatives $\partial \mathbf{T}_{\mathcal{FR}}/\partial \alpha_\iota$ of the image-to-image transformation w.r.t. the affine transformation parameters using expression (4); this involves 4×4 matrix manipulations only;
- computing the histogram $\partial \mathbf{H}/\partial \alpha_\iota$ as a linear combination of the histograms $\partial \mathbf{H}/\partial T_{\mathcal{FR},ij}$ and taking the sum of all entries of $\partial \mathbf{H}/\partial \alpha_\iota$ weighted by $\mathrm{d}I/\mathrm{d}h_{fr}$; this involves scanning the histogram $\mathbf{H}$ and all histograms $\partial \mathbf{H}/\partial \alpha_\iota$ once.

The computation of ∇I was implemented in C++. Because 12 histograms $\partial \mathbf{H}/\partial T_{\mathcal{FR},ij}$ need to be constructed, computation of ∇I is expected to be 12 times as expensive as computation of I itself. It was found experimentally that evaluation of ∇I is about 10 times more expensive than evaluation of I for the experiments discussed in Section 5. Computing a single component of ∇I is only slightly less complex, as most of the complexity is contained in constructing the histogram derivatives.

Figure 2 shows traces of I and $\partial I/\partial t_x$ for right-to-left translation of a CT and an MR image of the brain around the registered position. Note that although I itself is continuous, its gradient ∇I is not. A discontinuity in ∇I might appear each time the set of eight NNs $l_{k,m}$ in image $\mathcal{R}$ of any sample k in image $\mathcal{F}$ changes, if this results in other histogram entries $h_{f_k,r_{k,m}}$ being affected by k, whose derivative is weighted differently in Equation (8).

4. STRATEGIES FOR MAXIMIZATION OF MUTUAL INFORMATION

With the expressions for $I(\boldsymbol{\alpha})$ and $\nabla I(\boldsymbol{\alpha})$ derived in Section 3, various optimization strategies can be investigated for maximization of $I(\boldsymbol{\alpha})$. These are discussed briefly below for the general case of minimization of the N-dimensional function $f(\mathbf{x})$ with gradient $\nabla f(\mathbf{x})$. Note that maximizing $I(\boldsymbol{\alpha})$ corresponds to minimizing $I' = \log_2(\min(n_{\mathcal{F}}, n_{\mathcal{R}})) - I$ with $I' \geq 0$.

4.1. Powell's direction set method

Powell's direction set method only requires evaluations of f itself and not of ∇f. The method finds the N-dimensional minimum of f by repeatedly minimizing f in one dimension along a set of N different directions, each time starting from the minimum found in the previous direction using a one-dimensional (1-D) line minimization method such as Brent's (Press *et al.*, 1992). Powell's method incorporates a scheme to construct a set of conjugate directions iteratively. The set of directions is initialized with the basis vectors in each dimension in parameter space, but after each iteration in which all directions in the set are optimized over in turn, the overall distance moved in parameter space in that iteration is taken as a new direction. It can be shown that if f is a quadratic function, N mutually conjugate directions are obtained after N iterations, such that Powell's algorithm exactly minimizes a quadratic f in $N(N+1)$ line minimizations in all. For non-quadratic f, heuristics are needed to avoid the directions in the set becoming linearly dependent.

Our implementation closely follows the algorithm described in Press *et al.* (1992, pp. 412 ff.), but we reinitialize the direction set to the parameter basis vectors each time a new direction has been found and optimized over. Due to differences in image resolution in different directions and due to the specific shape of the objects in the scene, we found that the order in which the different parameters are considered and are optimized over influences optimization performance and registration robustness (Maes, 1998). For registration of images of the brain, optimization of the horizontal translations t_x and t_y and of the rotation ϕ_z around the vertical axis is better conditioned than optimization of the vertical translation t_z or of the pitch rotation ϕ_x around the right-to-left horizontal axis. We therefore optimize the parameters in each iteration in the order $(t_x, t_y, \phi_z, t_z, \phi_x, \phi_y)$.

4.2. Downhill simplex

This method, due to Nelder and Mead (Press *et al.*, 1992), does not require derivative information either and, unlike Powell's method, does not make use of a 1-D minimization algorithm. The method is initialized with $N+1$ points, defining a non-degenerate simplex in N-dimensional parameter space. This simplex is then deformed iteratively by reflection, expansion or contraction steps in order to move the vertices of the simplex towards the minimum of f. Convergence is declared when the fractional difference between the lowest and the highest function value evaluated at the vertices of the simplex is smaller than some threshold.

Our implementation of the simplex method is identical to its implementation given in Press *et al.* (1992, pp. 408 ff.). We initialize the simplex with the initial specified position in parameter space and with offsets around this position of $+5$ mm for translation and $+5^\circ$ for rotation in each of the parameter directions separately.

4.3. Steepest gradient descent

The steepest-gradient-descent method is the most straightforward method for incorporating gradient information in the minimization process. The minimum of the function is found by a number of consecutive 1-D line minimization steps using, for instance, Brent's algorithm, each step starting at the minimum found in the previous step and proceeding in the direction of the gradient at that point, i.e. the direction of steepest descent. Steepest gradient descent is in general not a very good algorithm. Because the gradient generally does not point to the optimum directly and because consecutive steps towards the optimum are necessarily at orthogonal angles, many tiny steps are usually required before reaching the optimum, especially when going down a long and narrow valley.

4.4. Conjugate-gradient methods

Conjugate-gradient methods try to overcome the problems associated with the steepest-gradient-descent approach by trying to construct the new direction as being conjugate to the previous one with respect to the function to minimize, rather than down the gradient direction. A scheme to construct a set of conjugate directions iteratively as the optimization proceeds has been proposed by Fletcher and Reeves (Press *et al.*, 1992). In each iteration i a line minimization is performed in the direction $\mathbf{d}_i$, starting at point $\mathbf{x}_i$ and leading to point $\mathbf{x}_{i+1}$. Initially, $\mathbf{d}_1$ is set to the gradient vector $\mathbf{g}_1$ at point $\mathbf{x}_1$. After each iteration a new direction $\mathbf{d}_{i+1}$ is constructed by

$$\gamma_i = \frac{\mathbf{g}_{i+1} \cdot \mathbf{g}_{i+1}}{\mathbf{g}_i \cdot \mathbf{g}_i}$$
$$\mathbf{d}_{i+1} = \mathbf{g}_{i+1} + \gamma_i \cdot \mathbf{d}_i.$$

It can be shown that if f is quadratic, the vectors $\mathbf{d}_i$ are mutually conjugate with respect to f (Press *et al.*, 1992), such that this scheme arrives at the exact minimum of f in N iterations, requiring N gradient evaluations in all. For non-quadratic f, more iterations are usually required.

Polak and Ribiere (Press *et al.*, 1992) introduced a small change to the Fletcher–Reeves algorithm using the form

$$\gamma_i = \frac{(\mathbf{g}_{i+1} - \mathbf{g}_i) \cdot \mathbf{g}_{i+1}}{\mathbf{g}_i \cdot \mathbf{g}_i}.$$

For quadratic f both expressions for γ_i are identical because then $\mathbf{g}_i \cdot \mathbf{g}_j = 0, i \neq j$, but for non-quadratic f there is some evidence that the Polak–Ribiere scheme converges faster than Fletcher–Reeves (Press *et al.*, 1992).

4.5. Quasi-Newton methods

The basic idea of variable metric or quasi-Newton methods is to build up iteratively a good approximation to the inverse Hessian matrix $\mathbf{J}^{-1}$ of the N-dimensional function f to be optimized. If the function can be assumed to be quadratic and $\mathbf{J}^{-1}$ is known, the step to take in iteration i from the current point $\mathbf{x}_i$ to the exact optimum $\mathbf{x}^*$ of f can be determined by setting $\nabla f = 0$ as in Newton's 1-D optimization method. This gives $\mathbf{x}^* = \mathbf{x}_i - \mathbf{J}^{-1} \cdot \nabla f(\mathbf{x}_i)$. Hence, $\mathbf{d}_i = -\mathbf{J}^{-1} \cdot \nabla f(\mathbf{x}_i)$ is taken as the direction in which a line minimization is performed in iteration i.

Initially, $\mathbf{d}_1$ is set to the gradient vector and the first approximation $\mathbf{J}_1^{-1}$ of $\mathbf{J}^{-1}$ is set to the identity matrix. Two approaches to iteratively update $\mathbf{J}_i^{-1}$ after each iteration using function and gradient evaluations have been proposed by Davidon–Fletcher–Powell (DFP) and by Broyden–Fletcher–Goldfarb–Shanno (BFGS) (Press *et al.*, 1992). For quadratic functions f, both schemes converge to $\mathbf{J}^{-1}$ in N iterations, hence requiring N gradient evaluations in all to arrive at the exact minimum of f. For non-quadratic f, BFGS has been recognized to be superior in details of round-off error and convergence tolerances (Press *et al.*, 1992).

4.6. Least-squares methods

Every minimization problem of a function $f(\mathbf{x})$ in N dimensions can be considered as a least-squares problem of recovering the N parameters $\mathbf{x}^*$ for which the least-squares figure of merit $f^2(\mathbf{x}^*)$ is minimal. An elegant method for general non-linear least squares has been put forth by Levenberg and Marquardt (Press *et al.*, 1992). The method varies smoothly between the steepest-descent and inverse-Hessian approaches by solving at each iteration i the incremental update $\delta\mathbf{x}_i$ from the current estimate $\mathbf{x}_i$ of the optimum to the next one from the equation

$$(\mathbf{J}_i + \lambda\mathbf{I}) \cdot \delta\mathbf{x}_i = -\nabla f^2 = -2 f \cdot \nabla f \tag{9}$$

with $\mathbf{J}_i$ the Hessian of f^2 at $\mathbf{x}_i$, $\mathbf{I}$ the identity matrix and λ a regularization parameter. For λ approaching zero, Equation (9) reduces to $\mathbf{J}_i \cdot \delta\mathbf{x}_i = -\nabla f^2$, which is the inverse-Hessian method, bringing us from $\mathbf{x}_i$ directly to the optimum if f^2 is a quadratic function. For sufficiently large values of λ, $\mathbf{J}_i + \lambda\mathbf{I}$ is diagonally dominant and Equation (9) reduces to $\delta\mathbf{x}_i \sim -\nabla f^2$, which is the steepest-descent method.

The Hessian matrix $\mathbf{J}_i = \{J_{kl}\}$ of f^2 is approximated by $J_{kl} \simeq 2\nabla f_k \cdot \nabla f_l$ by ignoring the second derivative terms. At each iteration, Equation (9) is solved for $\delta\mathbf{x}_i$, using a moderate value for λ, and f^2 is evaluated at $\mathbf{x}_i + \delta\mathbf{x}_i$. If $f^2(\mathbf{x}_i + \delta\mathbf{x}_i) \geq f^2(\mathbf{x}_i)$, λ is increased to favour the steepest-gradient approach and Equation (9) is solved for a new trial step $\delta\mathbf{x}_i$. If $f^2(\mathbf{x}_i + \delta\mathbf{x}_i) < f^2(\mathbf{x}_i)$, $\mathbf{x}_{i+1} = \mathbf{x}_i + \delta\mathbf{x}_i$ is taken as the new estimate for the optimum and a new iteration is started after having decreased λ to favour the inverse-Hessian approach. Convergence is declared when f^2 decreases by a negligible amount.

The Levenberg–Marquardt method has the advantage over the other gradient-based methods discussed herein that no line minimization is being performed in each iteration, which saves a lot of function evaluations.

4.7. Multiresolution techniques

All the above methods can be incorporated in a multiresolution scheme to increase speed performance. The optimization is performed first at lower resolution by using only a fraction of the voxels in image $\mathcal{F}$ to construct the joint histograms from which I and ∇I are computed. After convergence, the optimization proceeds at higher resolution, and eventually at full resolution, by taking more voxels of $\mathcal{F}$ into account.

A two-level multiresolution hierarchy can be constructed simply by subsampling the image $\mathcal{F}$ with integral sampling factors f_x, f_y and f_z along the x, y and z coordinate dimensions respectively using NN interpolation. This maintains geometric consistency and ensures that each sample in image $\mathcal{F}$ affects the same histogram bins at all resolution levels, such that the optimum is likely to be the same at all levels. Subsampling image $\mathcal{F}$ results in a speed-up by a factor of $F = f_x \cdot f_y \cdot f_z$ in the evaluation of I and ∇I. If the optimum

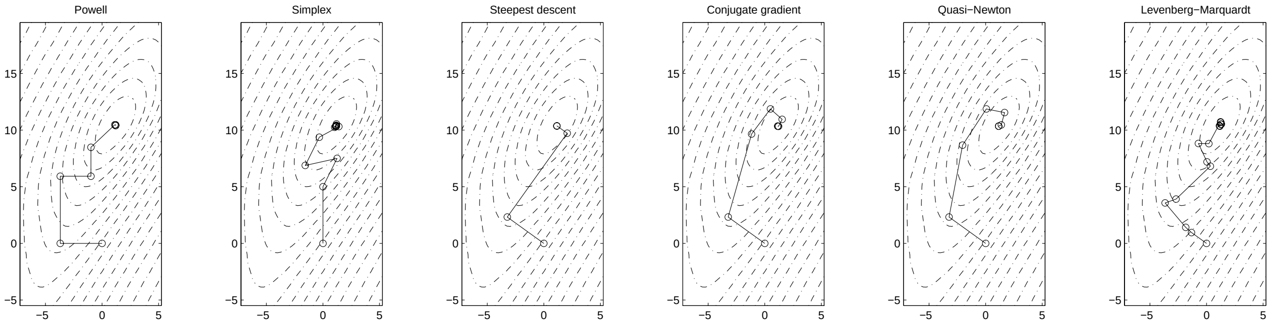


Figure 3. Optimization of mutual information using PV interpolation for a CT and an MR image of the head in a 2-D parameter subspace through the registration optimum, showing the paths in this plane followed by Powell, simplex, steepest gradient-descent, conjugate-gradient, quasi-Newton and Levenberg–Marquardt optimization methods. Horizontal axis, t_y, in millimetres; vertical axis, ϕ_x, in degrees.

converged to at lower resolution is close to the optimum at full resolution, it is to be expected that most of the evaluations of I and ∇I will be performed at lower resolution and that the number of evaluations at full resolution will be much smaller than when the optimization is done entirely at full resolution. The two-level multiresolution approach may thus be up to F times faster than the single-resolution approach.

However, the sampling factors f_x, f_y and f_z cannot be made arbitrarily large without introducing additional local optima and deteriorating registration robustness. The optimization at the lower-resolution level may then not converge to the optimum converged to at full resolution. If the optimization at the lower-resolution level does not provide a good starting position for the optimization at the higher-resolution level, the number of evaluations at full resolution will be about the same as for single-resolution optimization. For too large F, the multiresolution strategy might be a lot slower than single-resolution optimization, due to the additional, but useless, evaluations at lower resolution. Appropriate values for f_x, f_y and f_z can only be determined experimentally, although it is clear that this choice is related to the image resolution along each direction.

4.8. Example

Figure 3 shows traces in parameter space followed by the various optimization methods presented above starting from the same initial position for registration of a CT and an MR image of the brain. Both the CT and the MR image consist of axial slices with dimensions $256^2 \times 100$ and $256^2 \times 180$ and voxel sizes $0.94^2 \times 1.55$ and $0.98^2 \times 1.0$ respectively. The number of histogram bins was 100 for both images. The optimization of $I(\boldsymbol{\alpha})$ was constrained to a plane through the optimum of I by optimizing over two components of $\boldsymbol{\alpha}$, while keeping the other components constant and equal to their value at the optimum. The figures also show contour lines of I, which are almost elliptical curves centred around the optimum with major axes at angles of 45° to the parameter axes. Powell succeeds after two iterations at finding a new direction which leads almost directly to the optimum. Steepest descent, forced to proceed in directions that are each time orthogonal to the previous one, requires even fewer line minimizations than Powell in this 2-D example, but is nevertheless less efficient due to the expensive gradient computation. Conjugate-gradient and quasi-Newton behave fairly similarly, but again not better than Powell. Simplex and Levenberg–Marquardt follow rather erratic paths, but save a lot of function evaluations as no line minimizations are performed at each iteration. Simplex, requiring no gradient evaluations, is the most efficient in this case. Validation of the different optimization methods and a comparison of their performance for matching of CT and MR brain images is presented in Section 5.

5. VALIDATION

5.1. Experiments

The performance of the optimization methods discussed in Section 4 was evaluated for six-parameter rigid-body registration of CT and MR images of the brain using various multiresolution strategies. The experiments were performed on a set of nine patient data sets, consisting of CT, MR-MPRAGE, MR-PD, MR-T1 and MR-T2 images provided by J. M. Fitzpatrick as part of the RREP project (Fitzpatrick, 1994). These are not the data sets used in the original RREP project as reported in West *et al.* (1997), but a second series with higher-resolution images. The characteristics of these images are summarized in Table 1. MR-PD images were available for patients 1–4 only, while no MR-T2 image was available for patient 3. The MR-MPRAGE images of patients 1–4 are sagittal images, those of patients 5–9 are coronal

Table 1. Image characteristics

Image	Size	Voxels (mm)
CT	$512^2 \times (40–49)$	$(0.40–0.45)^2 \times 3.0$
MR-PD	$256^2 \times 52$	$(0.78–0.86)^2 \times 3.0$
MR-T1	$256^2 \times (51–52)$	$(0.78–0.86)^2 \times 3.0$
MR-T2	$256^2 \times 52$	$(0.78–0.86)^2 \times 3.0$
MR-MPRAGE	$256^2 \times 128$	$0.98^2 \times (1.25–1.66)$

Table 2. Subsampling factors $[\,f_x\, f_y\, f_z]$ and overall subsampling factor F applied to the columns, rows and slices respectively of the floating image at each level of the various multiresolution strategies evaluated in this study

Strategy	# Levels	Level 3	Level 2	Level 1
421	3	[441] (16)	[221] (4)	[111] (1)
441	2		[441] (16)	[111] (1)
332	2		[332] (18)	[111] (1)
331	2		[331] (9)	[111] (1)
222	2		[222] (8)	[111] (1)
221	2		[221] (4)	[111] (1)
111	1			[111] (1)

images, while all other images are axial. All images have a 12-bit intensity range.

For each patient data set, all MR images were registered to the CT image using the CT image as the reference image, resulting in 30 registration experiments in total, nine involving the high-resolution MR-MPRAGE images and 21 involving the somewhat lower-resolution MR-PD, MR-T1 and MR-T2 images. The number of histogram bins used to compute the mutual information criterion was 256 for both images in all experiments. All experiments were performed using Powell's direction set method (POW), downhill-simplex (SMP), steepest gradient-descent (STD), conjugate-gradient (CJG), quasi-Newton (QSN) and Levenberg–Marquardt (LVM) optimization methods at full image resolution and using each of five two-level and one three-level multiresolution strategies, resulting in a total of 1260 registration results. The subsampling factors used for each multiresolution strategy are summarized in Table 2. No smoothing of the images was applied prior to subsampling.

All optimizations for a given pair of images started from the same initial position with all six rotation and translation parameters set to zero, except for a rotational offset of 5° around one axis. The convergence parameters of the Brent line minimization algorithm used in POW, STD, CJG and QSN were identical for all methods: the absolute precision was set to 10^{-3}, the fractional precision to 10^{-2} and the maximum number of iterations to 10. The overall convergence criterion for each method was determined heuristically such that all methods yielded a similar precision. For POW, convergence was declared using the same criterion as in Maes *et al.* (1997) when the fractional decrease of the criterion in one iteration was smaller than 10^{-5}, while for the other methods a fractional tolerance of 10^{-6} was specified. Note that the convergence criterion is evaluated for POW after six line minimizations only, while for STD, CJG and QSN the convergence criterion is evaluated after every single line minimization and for SMP and LVM after every update of the trial solution. The same convergence criterion was used at each resolution level.

For each experiment the registration solution obtained using external markers as provided by the RREP project (Fitzpatrick, 1994) was used as a reference to evaluate registration accuracy. The error was evaluated as in Maes *et al.* (1997) by the mean norm of the difference vector between the reference and the computed transformation, evaluated at eight points near the brain surface.

All experiments were conducted on an SGI Octane workstation (IRIX 6.4, R10000 195 MHz, 17 SPECfp95). Because the load of this machine varied while the experiments were done, it is unreliable to compare different methods by the CPU time required to reach convergence. Instead, different experiments are compared by their number of equivalent full-resolution criterion evaluations N_e, defined as the weighted sum of the number of function and gradient evaluations at each resolution level, taking into account the subsampling factors at each level and the relative difference in complexity between function and gradient evaluations:

$$N_e = \sum_{r=1}^{R} \frac{N_{f,r} + \eta \cdot N_{g,r}}{F_r}$$

with R the number of resolution levels (one, two or three in our experiments), $N_{f,r}$ and $N_{g,r}$ the number of function and gradient evaluations at resolution level r, F_r the overall subsampling factor at that level and η the relative complexity of the gradient evaluation compared with that of the function evaluation. We used $\eta = 12$, which is rather conservative because a factor of 10 was found experimentally as discussed in Subsection 3.3. The values of N_e were aggregated over all patients for each method and each multiresolution strategy for the higher-resolution MR-MPRAGE images in one group and for the lower-resolution MR-PD, MR-T1 and MR-T2 in another. The performance of the various methods was evaluated by comparing their mean N_e values over all 30 registration experiments using the one-sided t-test with a significance level of $p = 0.005$.

5.2. Results

5.2.1. *Accuracy*

The recovered registration parameters ranged from -29 to $+13$ mm for translation and from -28° to $+14^\circ$ for rotation. The accuracy of all 1260 experiments with respect to the external-marker-based reference solution ranged from 0.5 to 3.1 mm with a mean of 1.78 mm. All errors were thus <1 CT voxel (Maes, 1998). It was found that the error is independent of the optimization method that was used: for each of the 30 different registration experiments, error variations among the 42 different optimization results was $<5\%$, indicating that each of these converged to the same optimum of MI.

5.2.2. *Precision*

The variation in the optimum found by each method was evaluated by comparing for each of the 30 experiments the registration parameters $\boldsymbol{\alpha}$ for each of the 42 optimization results with the parameters found for the best optimization result $\boldsymbol{\alpha}^*$, i.e. the one for which the mutual information criterion was maximal. The parameter vectors were compared by the Euclidean norm $|\delta\boldsymbol{\alpha}| = |\boldsymbol{\alpha} - \boldsymbol{\alpha}^*|$ of their difference vector.

The mean values of $|\delta\boldsymbol{\alpha}|$ over all experiments are 0.017, 0.028, 0.041, 0.030, 0.040 and 0.034 for POW, SMP, STD, CJG, QSN and LVM respectively, while the maximal values are 0.153, 0.170, 0.161, 0.142, 0.141 and 0.166 respectively. POW is on average significantly more precise than any other method ($p < 0.05$) and in 23 out of 30 registration experiments POW yields the best optimum, which can be explained by the fact that the stop criterion is verified after only six consecutive line minimizations. With the same fractional tolerance specified on the registration criterion, SMP and CJG are on average significantly more precise than STD, QSN and LVM; LVM is on average significantly more precise than STD and QSN. But the maximal values of $|\delta\boldsymbol{\alpha}|$ are almost identical for all methods, such that their worst-case precision is approximately the same. All values of $|\delta\boldsymbol{\alpha}|$ are smaller than 0.2, indicating that for each experiment none of the optimization results differ by more than 0.2 mm for the translational and 0.2° for the rotational parameters. All differences between corresponding results are therefore <1 CT voxel everywhere in the image and all methods always achieve subvoxel precision for the specified convergence parameters.

When comparing the precision between different multiresolution strategies, no significant differences were observed except for the three-level [421] multiresolution strategy, which was found to be on average significantly more precise than all other strategies, but for the gradient-based optimization methods only.

5.2.3. *Performance*

The performance of each of the optimization strategies is summarized in Figure 4. This figure shows box and whisker plots (Matlab, 1996) for the values of N_e obtained for all 30 CT/MR experiments with each of the six optimization methods and each of the seven multiresolution strategies. The results were aggregated for the CT/MR-MPRAGE and the CT/MR-(PD, T1, T2) experiments as no significant differences were observed between both groups. The mean, minimum and maximal values of N_e for each case and of the recorded CPU time are tabulated in Tables 3 and 4.

The plots show that the use of any of the two- or three-level multiresolution approaches results in a decrease of computational complexity by a factor of two for POW and 3–5 for the other methods. For all methods, multiresolution strategies [222], [331], [332], [441] and [421] perform significantly better than [111] and also significantly better than [221] for SMP, STD, CJG and QSN. However, subsampling factors >8 at the lower-resolution level do not offer any significant speed advantage: for all methods there is no significant difference in performance between the multiresolution approaches [222], [331], [332] and [441]. This can be explained by the fact that the computational complexity is mainly determined by the number of iterations and gradient evaluations at full resolution, while optimization at lower-resolution levels has only a small impact on the overall complexity. The distance between the optima converged to at lower and at full resolution increases for larger subsampling factors, such that more full-resolution evaluations are required to reach convergence. Hence, the advantage of faster evaluations at lower resolution is cancelled by the additional number of evaluations required at full resolution. The three-level strategy [421] performs worse than the two-level strategies [222], [331], [332] and [441] for CJG, but better than [441] for SMP and no significant differences were observed for the other methods.

When comparing the different optimization methods applied at full resolution [111], POW performs not significantly worse than any other method and performs significantly better than STD and QSN at [111] and [221]. STD and QSN are not faster than any other method for any multiresolution strategy. SMP, CJG and LVM perform significantly better than POW, STD and QSN at any of the two- or three-level multiresolution strategies, while no other method is faster than SMP, CJG or LVM. No significant differences in performance have been observed between SMP, CJG and LVM, except for SMP and LVM being faster than CJG at [421], but SMP being slower than CJG at [441]. The average speed-up factors that have been observed with each

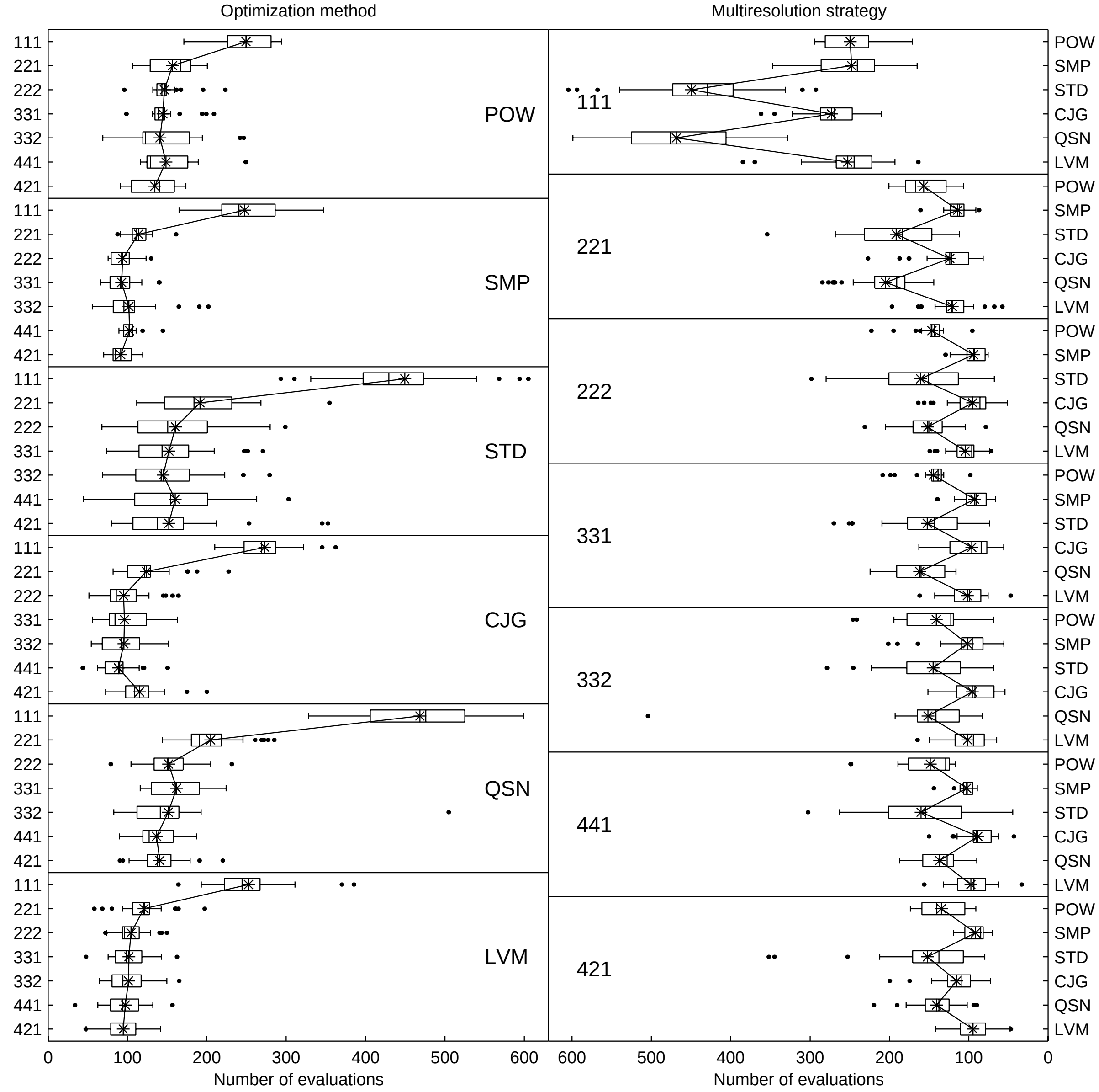


Figure 4. Box and whisker plots of the number of equivalent function evaluations N_e for all 30 CT/MR registration experiments for each optimization method and each multiresolution strategy. Each box has lines at the lower quartile, median and upper quartile values. The whiskers are lines extending from each end of the box to show the extent of the rest of the data. The length of the whiskers is at most equal to the inter-quartile range. Outliers with values beyond the ends of the whiskers are indicated by '•'. The mean value is indicated by '∗'. The plots on the left and on the right show the same information, grouped by optimization method and by multiresolution strategy respectively.

method and its best multiresolution strategy compared with POW applied at [111] and computed from the N_e values are tabulated in Table 5. The best methods are SMP at [421], CJG at [441] and LVM at [421], which systematically perform three times better than POW at full resolution. In individual cases, these multiresolution approaches may be up to six times faster than POW at full resolution. The CPU times in Table 4 show that the registration of the CT and MR-MPRAGE images is performed in 5 min CPU time on average using SMP, CJG or LVM and the [441] multiresolution

Table 3. Mean number of equivalent full-resolution function evaluations for each method and multiresolution strategy over all 30 CT/MR registration experiments. The numbers in parentheses are the minimal and maximal values for each case. Mean values that are significantly lower than the corresponding values obtained with POW using the same multiresolution strategy are indicated with * for each method.

	POW		SMP		STD		CJG		QSN		LVM	
421	134	(91, 174)	91*	(70, 119)	152	(80, 352)	115*	(73, 200)	141	(90, 220)	95*	(47, 142)
441	148	(116, 249)	103*	(89, 144)	160	(45, 303)	89*	(43, 150)	137	(90, 187)	97*	(34, 156)
332	141	(69, 246)	102*	(56, 202)	145	(69, 279)	96*	(54, 152)	151	(83, 504)	101*	(65, 165)
331	145	(98, 209)	93*	(66, 140)	152	(74, 270)	96*	(56, 163)	162	(116, 224)	102*	(47, 162)
222	147	(96, 223)	94*	(76, 130)	161	(68, 299)	95*	(52, 164)	152	(79, 231)	104*	(72, 149)
221	157	(107, 201)	114*	(87, 161)	191	(112, 354)	124*	(82, 227)	205	(144, 285)	121*	(58, 197)
111	249	(171, 294)	248	(165, 347)	449	(293, 905)	273	(210, 362)	468	(328, 656)	252	(164, 385)

Table 4. Mean CPU time in seconds for each method and multiresolution strategy over all patient data sets for the nine CT/MR-MPRAGE experiments (top) and the 21 CT/MR-(T1, T2, PD) experiments (bottom). The numbers in parentheses are the minimal and maximal values for each case.

	POW		SMP		STD		CJG		QSN		LVM	
421	417	(247, 637)	322	(178, 466)	470	(289, 594)	407	(292, 521)	509	(300, 725)	314	(203, 429)
441	445	(249, 795)	351	(189, 532)	567	(134, 949)	339	(220, 573)	482	(271, 752)	323	(113, 471)
332	403	(200, 607)	363	(165, 555)	519	(315, 875)	342	(220, 523)	516	(316, 818)	399	(294, 522)
331	460	(283, 670)	301	(153, 382)	561	(248, 854)	348	(207, 499)	610	(337, 812)	407	(271, 500)
222	465	(296, 714)	314	(166, 442)	505	(188, 735)	311	(168, 449)	510	(356, 682)	407	(271, 500)
221	481	(312, 673)	421	(204, 581)	698	(337, 1086)	405	(255, 569)	765	(515, 1017)	454	(235, 654)
111	764	(514, 1216)	789	(390, 1086)	1602	(910, 2597)	961	(663, 1120)	1615	(975, 2146)	957	(620, 1283)
421	252	(186, 334)	172	(136, 221)	391	(220, 845)	282	(167, 503)	362	(236, 559)	263	(170, 389)
441	255	(178, 400)	177	(134, 218)	375	(173, 724)	195	(135, 297)	335	(232, 481)	267	(188, 424)
332	241	(106, 381)	167	(91, 307)	327	(149, 572)	213	(134, 359)	360	(217, 1057)	254	(173, 399)
331	248	(196, 338)	175	(123, 272)	356	(188, 621)	219	(147, 378)	379	(259, 534)	259	(134, 444)
222	241	(196, 387)	165	(124, 241)	384	(195, 709)	220	(124, 396)	354	(173, 659)	262	(172, 404)
221	273	(180, 340)	205	(155, 251)	441	(245, 646)	286	(201, 535)	469	(354, 659)	307	(184, 432)
111	367	(251, 468)	380	(244, 564)	927	(634, 1399)	543	(426, 705)	990	(632, 1282)	605	(464, 818)

strategy, while about 13 min CPU time is required on average using POW at full resolution.

6. DISCUSSION

Expressions for the gradient of MI were derived analytically by differentiation of the joint histogram with respect to the registration parameters. The histogram is a continuous function of the registration parameters when PV interpolation is used, which allows us to compute the gradient exactly. The gradient is effectively zero at the optimum and no heuristics are needed to obtain a reliable estimate of the gradient, for instance by using finite differences.

The performance of the various optimization strategies was compared by the number of equivalent full-resolution function evaluations required to reach convergence. This assumes that equivalent convergence criteria were specified for each method and that all methods have a similar precision. By comparing the various optimization results obtained with different optimization methods for the same registration experiment, we found that POW most often yields the best optimum, which can be explained by the fact that the stop criterion is evaluated after only six line minimizations. Although the worst-case precision was the same for all methods, the STD and QSN methods on average have a somewhat worse precision than the other methods, which indicates that these methods are more likely to generate poor optimization directions in which the registration criterion only slightly changes.

SMP, CJG and LVM were found to be the most performant methods for any of the two- or three-level multiresolution strategies, systematically outperforming the other methods by 30–80%. POW is as efficient as SMP, CJG and LVM for single-resolution optimization, but less efficient when

Table 5. Best average and maximal speed-up factors and multiresolution strategies for each method compared to POW at full resolution.

Method	Speed-up factor	Strategy
POW	1.9 (3.6)	[331]
SMP	2.8 (4.0)	[421]
STD	1.9 (4.0)	[331]
CJG	3.0 (6.3)	[441]
QSN	1.9 (2.8)	[441]
LVM	2.8 (5.8)	[421]

multiple resolution levels are used, which indicates that SMP, CJG and LVM better exploit the advantage of a good starting position computed at a lower-resolution level. SMP performs as well as CJG and LVM, but is much easier to implement as it does not require evaluation of the gradient expressions derived in Section 3. SMP is therefore to be preferred for multiresolution optimization, while for single resolution POW is our method of choice, as it is the most precise.

It was found that large subsampling factors ([332], [441]) do not result in significantly better performance compared with moderate subsampling factors ([222], [331]), because the performance is dominated by the number of function and gradient evaluations required at the final full-resolution level. Much larger speed-ups may be gained by not including the full-resolution level in the multiresolution hierarchy and simply stopping the optimization when convergence is reached at lower resolution, trading precision for speed. This was investigated for the LVM method. Restricting the optimization for the two-level multiresolution strategies to the lower-resolution level only decreases the precision when compared with the solution found when including full resolution to 0.25 mm and 0.25° on average, but differences larger than 0.5 mm or 0.5° were observed in some cases which is no longer subvoxel. However, the performance was increased to a speed-up of more than 10 for the [441] and [332] multiresolution strategies compared with the same method applied at full resolution. Imposing a more stringent stop criterion may increase precision while still keeping a significant speed advantage.

The use of three rather than two multiresolution levels does not increase the speed performance, due to the additional number of function and gradient evaluations that is required to reach convergence at the intermediate resolution level. If the optimization at the selected lowest-resolution level already converges to an optimum which is close to the optimum at full resolution, it is not efficient to first do another optimization at a somewhat higher resolution instead of considering full resolution right away. However, using more than one subsampled resolution level might be appropriate if the optimization at the selected lowest-resolution level fails. Especially for lower-resolution images and for large subsampling factors, subsampling may introduce additional local minima in the registration criterion such that the optimization at the lower-resolution level may converge to an optimum that is still far off from the one at full resolution. This deteriorates speed performance, as many more evaluations at full resolution will be required. In the experiments discussed above this occurred a few times for the STD and QSN methods, which can be seen from the outliers in Figure 4.

We have not experimented with specifying different convergence parameters for the lower and the full-resolution level. Pluim (1996) investigated the influence of accepting lower precision at the lower-resolution level by increasing the fractional tolerance for this level, but found no significant improvement in registration performance. This can be explained by the fact that the performance is mainly determined by the number of evaluations at full resolution and that the additional number of evaluations required at lower resolution to reach a higher precision at that level contributes only marginally to the overall computational cost.

The various multiresolution levels were created by subsampling the floating image using integral subsampling factors and nearest-neighbour interpolation, such that no new intensity values were introduced by the sampling process. It is unclear how the use of non-integral subsampling factors with trilinear interpolation, prior smoothing of the images or reducing the number of image histogram bins would influence the behaviour of the registration criterion at the lower-resolution level. Pluim (1996) has experimented with various multiresolution hierarchies using Powell's algorithm and reported better results using the original intensity values as we did than when first smoothing the images prior to downsampling. This may be explained by the fact that the smoothing may change the shape of the iso-intensity objects on which the registration is based such that the boundaries of these objects appear different at lower than at full resolution, which induces differences and inconsistencies in the optimal registration position at different resolution levels. This is avoided by the subsampling scheme applied above.

The optimization methods evaluated here are all local optimization methods. Global optimization methods, such as simulated annealing, have not been considered. These methods aim to find the global optimum in the presence of many local optima of comparable strength. However, the optimum that is being searched for when maximizing mutual information is not the global optimum, but rather the local optimum within some capture range. Simulated annealing or other global optimization methods therefore do not offer specific advantages over the methods considered here.

All experiments discussed in Section 5 involve matching of high-resolution CT and MR images of the brain. The results cannot be simply generalized to lower-resolution data, especially not for the gradient-based methods, because of the gradient being noisier if fewer samples are available. However, registration of lower-resolution images is less computationally expensive and speed performance is therefore less of an issue. Powell's method applied at full image resolution has been demonstrated to be very robust and sufficiently performant for matching lower-resolution images (Maes *et al.*, 1997; West *et al.*, 1997). Experiments on CT and MR images of the prostate (512 × 512 matrix, 0.65 mm pixel size, 5 mm slice distance, 22 MR slices, 40 CT slices) have confirmed the above results.

7. CONCLUSION

The mutual information registration criterion is a continuous function of the affine registration parameters when PV interpolation is being used. In this paper we have evaluated the performance of various gradient- and non-gradient-based optimization strategies, using analytical and exact expressions for the gradient of the registration criterion. Although the affine gradient computation is a computationally expensive operation, large speed-ups of the registration process can be achieved without a loss of robustness by embedding the optimization in a multiresolution scheme. A two-level multiresolution approach using the simplex, conjugate-gradient or Levenberg–Marquardt optimization method was found to be the most efficient, systematically outperforming Powell's method applied at full image resolution by a factor of at least three. The results reported in this paper demonstrate that high-resolution CT and MR images can be robustly matched in $<$5 min CPU time on current workstations. This is comparable with the time that is usually required to transfer the images over the local hospital network from the scanners to the workstation, to load the images from disk into computer memory and to display them on the workstation screen. It is to be expected that the availability of more performant registration algorithms will further stimulate the use of multimodal data in many applications in routine clinical practice.

ACKNOWLEDGEMENTS

This research was supported by IBM Belgium (Academic Joint Study) and by the National Fund for Scientific Research, Belgium, under grant numbers FGWO 3.0115.92, 9.0033.93 and G.3115.92. This research is performed within the context of the EASI project 'European Applications for Surgical Interventions', supported by the European Commission under contract HC1012 in their '4th Framework Telematics Applications for Health' RTD programme. The partners in the EASI consortium are Philips Medical Systems, Nederland BV, Philips Research Laboratory Hamburg, the Center for Processing Speech and Images of the Katholieke Universiteit Leuven, the Image Sciences Institute of Utrecht University and University Hospital Utrecht, The National Hospital for Neurology and Neurosurgery in London, and the Computational Imaging Science group of Guy's, King's and St Thomas' Hospitals Medical Schools, London. The images and the standard transformations were provided as part of the project 'Evaluation of Retrospective Image Registration', National Institutes of Health, Project Number 1 R01 NS33926-01, Principal Investigator, J. Michael Fitzpatrick, Vanderbilt University, Nashville, TN.

REFERENCES

Collignon, A., Maes, F., Vandermeulen, D., Suetens, P. and Marchal, G. (1995) Automated multi-modality image registration based on information theory. In Bizais, Y., Barillot, C. and di Paola, R. (eds), *Proc. XIVth Int. Conf. Information Processing in Medical Imaging, Computational Imaging and Vision*, Vol 3, pp. 263–274, Ile de Berder, France. Kluwer Academic Publishers, Dordrecht.

Duda, R. O. and Hart, P. E. (1973) *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York.

Fitzpatrick, J. M. (1994) *Evaluation of Retrospective Image Registration, NIH Project Nr. 1 R01 NS33926-01*. Vanderbilt University, Nashville, TN.

Maes, F. (1998) *Segmentation and Registraion of Multimodal Medical Images: from Theory, Implementation and Validation to a Useful Tool in Clinical Practice*. Ph.D. Thesis, K. U. Leuven, Faculteit Toegepaste Wetenschappen, Leuven, Belgium.

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G. and Suetens, P. (1997) Multi-modality image registration by maximization of mutual information. *IEEE Trans. Med. Imag.*, 16, 187–198.

Matlab (1996) *Matlab Statistics Toolbox User's Guide*. The MathWorks, Inc., 24 Prime Park Way, Natick, MA.

Meyer, C., Boes, J. L., Kim, B., Bland, P. H., Wahl, R. L., Zasadny, K. R., Kison, P. V., Koral, K. and Frey, K. A. (1997) Demonstration of accuracy and clinical versatility of mutual information for automatic multimodality image fusion using affine and thin plate spline-warped geometric deformations. *Med. Image Anal.*, 1, 195–206.

Pluim, J. P. W. (1996) *Multi-modality Matching Using Mutual Information.* Master's thesis, University of Groningen, Department of Computing Science, Groningen, The Netherlands.

Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T. (1992) *Numerical Recipes in C*, second edition. Cambridge University Press, Cambridge.

Studholme, C., Hill, D. L. G. and Hawkes, D. J. (1996) Automated 3-D registration of MR and CT images of the head. *Med. Image Anal.*, 1, 163–175.

Viola, P. and Wells, W. M. III (1995) Alignment by maximization of mutual information. In *Proc. 5th Int. Conf. Computer Vision*, pp. 16–23, Cambridge, MA.

Wells W. M. III, Viola, P., Atsumi, H., Nakajima, S. and Kikinis, R. (1996) Multi-modal volume registration by maximization of mutual information. *Med. Image Anal.*, 1, 35–51.

West, J. *et al.* (1997) Comparison and evaluation of retrospective intermodality brain image registration techniques. *J. Comp. Assis. Tomogr.*, 21, 554–566.