IEEE Transactions on Medical Imaging

Accelerated Nonrigid Intensity-Based Image Registration Using Importance Sampling

Journal:	*Transactions on Medical Imaging*
Manuscript ID:	TMI-2008-0617.R1
Manuscript Type:	Full Paper
Date Submitted by the Author:	31-Dec-2008
Complete List of Authors:	Bhagalia, Roshni; University of Michigan, Electrical Engineering and Computer Science Fessler, Jeffrey; University of Michigan, Electrical Engineering and Computer Science Kim, Boklye; University of Michigan, Radiology
Keywords:	gradient optimization, intensity-based registration, importance sampling, stochastic approximation

Accelerated Nonrigid Intensity-Based Image Registration Using Importance Sampling

Roshni Bhagalia, Jeffrey A. Fessler, *Fellow, IEEE,* and Boklye Kim

Abstract

Nonrigid image registration methods using intensity-based similarity metrics are becoming increasingly common tools to estimate many types of deformations. Nonrigid warps can be very flexible with a large number of parameters and gradient optimization schemes are widely used to estimate them. However for large datasets, the computation of the gradient of the similarity metric with respect to these many parameters becomes very time consuming. Using a small random subset of image voxels to approximate the gradient can reduce computation time. This work focuses on the use of importance sampling to reduce the variance of this gradient approximation. The proposed importance sampling framework is based on an edge-dependent adaptive sampling distribution designed for use with intensity-based registration algorithms. We compare the performance of registration based on stochastic approximations with and without importance sampling to that using deterministic gradient descent. Empirical results, on simulated MR brain data and real CT inhale-exhale lung data from 8 subjects, show that a combination of stochastic approximation methods and importance sampling accelerates the registration process while preserving accuracy.

This work was supported in part by grants NIH8RO1EB003O9 and 1PO1CA87634.

R. Bhagalia (e-mail: rbhagali@umich.edu) and J. A. Fessler (e-mail: fessler@umich.edu) are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan 48109 USA.

B. Kim (e-mail: boklyek@umich.edu) is with the Department of Radiology, University of Michigan Medical School, Ann Arbor, MI 48109 USA.

Accelerated Nonrigid Intensity-Based Image Registration Using Importance Sampling

I. Introduction

NONRIGID registration algorithms estimate a warp or deformation with many ($\gg 12$ (3D affine)) degrees of freedom that appropriately maps one image onto another. The estimated warp models can be either parametric [1]–[4] or non-parametric [5], [6]. In this paper, we focus on intensity-based image registration methods that estimate parameterized warp models by solving an optimization problem:

$$\hat{\theta} = \arg\max_{\theta} \Psi(\theta); \tag{1}$$

where Ψ is the similarity metric and $\hat{\theta}$ is the estimate of the p dimensional vector of warp parameters.

In registration scenarios that use differentiable intensity-based similarity metrics and gradient optimization methods, it is possible to derive an analytical expression for the gradient of the similarity metric $\nabla_{\theta}\Psi(\theta)$. However for large image datasets, the large number of warp parameters in most nonrigid registration methods makes the gradient calculation time consuming. A simple strategy to reduce this computation time is to use a small random subset of image voxels to approximate the gradient [7].

Since this randomization of the gradient in effect makes the search direction a random variable, these techniques cannot be used with algorithms like Conjugate Gradients that endeavor to maintain the conjugacy of successive search directions. Furthermore while it is possible to approximate the Hessian, because the random sample-size is small, its accuracy is suspect. Hence step-sizes based on the inverse of the Hessian, as in the Levenberg-Marquardt scheme, may not be reliable. It was reported in [7] that an analytical gradient-based optimizer [2], [3], using a random sub-sampling technique to approximate the gradient, performed better than that using gradient approximations based on finite differences [8] and simultaneous perturbation [9].

The speed and accuracy of such registration algorithms depend on the quality of the gradient approximation obtained via random sampling. The subset of random voxel locations is typically drawn using uniform sampling (US). Here we present an alternative data-driven, non-uniform sampling strategy that can be used efficiently to improve these gradient approximations. We argue that image edges strongly influence intensity-based registration estimates. Consequently, we propose the use of importance sampling (IS) based on a sampling distribution that emphasizes image edges to improve the gradient approximations.

Section II-A casts image registration in a Stochastic Approximation framework. Importance sampling is described in Sec. II-B; a non-uniform sampling distribution for intensity-based registration is developed in Sec. II-C; and an efficient implementation strategy is outlined in Sec. II-E. Sec. III uses simulated 3D MRI volumes to compare the performance of multi-modal image registration using both IS and US with that using a deterministic gradient descent

optimizer. Lastly we demonstrate the application of IS to register real inhale-exhale lung CT data using deformable B-spline warps. The quality of the registration for CT data is quantified using expert identified landmarks. These results suggest that IS based on the sampling distribution designed in this work can accelerate intensity-based nonrigid registration algorithms while preserving accuracy.

II. Theory

A. Stochastic Approximation

Image registration based on random sampling becomes a stochastic approximation technique with the following updates:

$$\theta_{k+1} = \theta_k + a_k \hat{g}(\theta_k); \tag{2}$$

where θ_k is the warp parameter estimate at the kth iteration, $\hat{g}(\theta_k)$ is an approximation of the gradient $\nabla_\theta \Psi(\theta)$ at θ_k and a_k is the step-size. The iterative updates given by (2) require only an approximation of the gradient $\nabla_\theta \Psi(\theta)$; the similarity metric $\Psi(\theta)$ itself does not need to be computed. Stochastic approximation (SA) is used to find the zeros of a function (here $\nabla_\theta \Psi(\theta)$) when only noisy function evaluations are available [8], [10]. SA methods aim to find the unknown zeros by successively reducing the inaccuracy in their estimates. They have been applied successfully to numerous applications in the fields of statistical modeling and controls. In gradient-based image registration, SA techniques can be used to estimate warp parameters that maximize the similarity metric by steadily reducing the imprecision introduced in successive gradient approximations.

A now common SA approach was first introduced by Robbins and Monro [11]. This method aims to reduce the inaccuracy in its estimates by gradually reducing the step-size of the iterations; for brevity we call this technique Step-SA. Step-SA requires that the number of points (image voxels) used to approximate the gradient, i.e., the sample-size, remains fixed over iterations. The step-size sequence, designed to guarantee convergence of the optimizer, is a non-increasing non-zero sequence $\{a_k\}, k \in \mathbb{N}$ such that $\sum_{k=1}^{\infty} a_k = \infty$ and $\sum_{k=1}^{\infty} {a_k}^2 < \infty$. Clearly there are numerous sequences that describe a valid step-size progression. In practice the step-size sequence is chosen heuristically for a given application.

Unlike Step-SA, sample-size controlled SA (Samp-SA) [12] keeps the step-size constant. Errors in parameter estimates are reduced by progressively increasing the sample-size used to approximate the gradient. The slowest sample-size growth rate that ensures convergence is proportional to $\ln(k)$ where k is the iteration number [12]. Using a slow growth rate should reduce computation time. Both techniques effectively average out the approximation error as the iterations progress, yielding convergence.

Irrespective of the SA scheme used, the efficiency of these methods for image registration applications depends on the bias and variance properties of the underlying gradient approximation based on a small random subset of image voxels. This work focuses on the use of importance sampling to enhance the performance of registration algorithms by reducing the variance of such gradient approximations without introducing any bias. Since we use SA iterations given by (2), we restrict our attention to the bias and variance properties of the gradient approximation $\hat{g}(\theta)$ alone.

The similarity metric $\Psi(\theta)$ need not be computed or approximated. In the following section we briefly review importance sampling and identify image regions that strongly influence intensity-based registration. Subsequently we describe an appropriate adaptive sampling distribution that emphasizes samples from these regions. Further, a simple strategy to efficiently implement the sampling distribution is discussed.

B. Importance Sampling

Importance sampling (IS) is a variance reduction technique capable of incorporating knowledge of the quantity being approximated into the sampling process. IS recognizes that certain types of random samples can affect the approximation more than others and utilizes a sampling distribution that emphasizes these important samples. Such a biased distribution would produce a biased estimator; however by weighting the samples appropriately this bias can be preempted. For completeness we briefly summarize IS along the lines of [13]. To study the variance reduction afforded by IS, consider estimating a computationally intractable integral $\Phi = \int_\Omega f(x)dx$. This integral can be expressed as the expectation of a (non-linear) function of a uniformly distributed random vector U such that,

$$\begin{aligned} \Phi = \int_\Omega f(x)dx &= |\Omega| \int_\Omega f(u)\frac{1}{|\Omega|}du \\ &= |\Omega| E_U(f(U)),\ U \backsim \text{Unif}_\Omega \end{aligned} \tag{3}$$

where, Unif_Ω is the uniform distribution over Ω given by

$$\text{Unif}_\Omega(u) = \begin{cases} \frac{1}{|\Omega|}, & u \in \Omega \\ 0 & \text{else.} \end{cases}$$

Alternatively, the intractable integral Φ can also be written as the expectation of a function of a *non-uniform* random variable Y, given by:

$$\begin{aligned} \int_\Omega f(x)dx &= |\Omega| \int_\Omega \frac{f(y)}{w(y)}\frac{w(y)}{|\Omega|}dy \\ &= |\Omega| E_Y\left(\frac{f(Y)}{w(Y)}\right),\ Y \backsim \hat{P}_Y, \end{aligned} \tag{4}$$

where, the non-uniform distribution $\hat{P}_Y$ is given by

$$\hat{P}_Y(y) = \begin{cases} \frac{w(y)}{|\Omega|}, & y \in \Omega \\ 0 & \text{else.} \end{cases}$$

To gain any advantage by using $E_Y(.)$ over $E_U(.)$, the function $w(y)$ should be chosen carefully.

In practice, the expectations above are approximated by their sample means using N i.i.d. samples of random vectors $U_n \backsim \text{Unif}_\Omega$ and $Y_n \backsim \hat{P}_Y$. Ignoring the proportionality constant $|\Omega|$, consider the following estimates of the integral in (3);

$$\begin{aligned} \hat{\Phi}_{\text{uni}} &\triangleq \frac{1}{N}\sum_{n=1}^{N} f(U_n) \approx E_U(f(U)) \\ \hat{\Phi}_{\text{imp}} &\triangleq \frac{1}{N}\sum_{n=1}^{N} \frac{f(Y_n)}{w(Y_n)} \approx E_Y\left(\frac{f(Y)}{w(Y)}\right). \end{aligned}$$

$\hat{\Phi}_{\text{uni}}$ corresponds to the uniform sampling (US) case and $\hat{\Phi}_{\text{imp}}$ is the estimate obtained by importance sampling (IS). Both $\hat{\Phi}_{\text{uni}}$ and $\hat{\Phi}_{\text{imp}}$ are unbiased with expectations proportional to the original integral in (3). Since the random samples are i.i.d., the variances of the two estimates are given by

$$\text{var}(\hat{\Phi}_{\text{uni}}) = \frac{1}{N}\text{var}(f(U)) \text{ and } \text{var}(\hat{\Phi}_{\text{imp}}) = \frac{1}{N}\text{var}\left(\frac{f(Y)}{w(Y)}\right).$$

IS based on the sampling distribution $\hat{P}_Y$ is beneficial only if $\hat{P}_Y(y) = \frac{w(y)}{|\Omega|}$ is chosen to ensure that $\text{var}(\hat{\Phi}_{\text{imp}}) \ll \text{var}(\hat{\Phi}_{\text{uni}})$. This is possible if and only if the function $\frac{f(.)}{w(.)}$ has lower variance than $f(.)$ alone. Thus the weights $w(.)$ and correspondingly the sampling distribution $\hat{P}_Y$ should be chosen to be similar in shape to the original integrand $f(.)$, ensuring that the function $\frac{f(.)}{w(.)}$ is approximately constant.

C. Sampling Distributions for Image Registration

To design a meaningful sampling distribution for gradient-based image registration, we first identify image regions that contribute significantly to the gradient of the similarity metric. Consider registration between a pair of intensity images, namely the reference image with N_u voxels and the homologous image with N_v voxels. These images are assumed to be sets of samples $\tilde{u}_i = u(x_i)$, $i = 1, 2, \ldots N_u$, and $\tilde{v}_j = v(y_j)$, $j = 1, 2, \ldots N_v$, of continuous intensity functions $u(.)$ and $v(.)$ respectively. These continuous functions are sampled at coordinates $x_i \in \mathbb{R}^3$ and $y_j \in \mathbb{R}^3$ respectively.

Most nonrigid registration algorithms assume that image coordinates are related by a warp $T_{\theta_*} : \mathbb{R}^3 \rightarrow \mathbb{R}^3$. The vector of unknown warp parameters $\theta_* \in \mathbb{R}^p$ is estimated iteratively by the algorithm. At each iteration, the current estimate $\theta = \theta_k$ is used to find intensities at coordinates $\{y_i^\theta = T_\theta(x_i)\}_{i=1}^{N_u}$ in the homologous image corresponding to each reference voxel location. These transformed coordinates rarely lie on the sampling grid points and hence their corresponding intensity values $\{\hat{v}_i^\theta \approx v(y_i^\theta)\}$ are not known. Intensity-based similarity metrics commonly approximate these unknown intensities by modeling the continuous intensity function $v(.)$ using an appropriate interpolation kernel. Specifically, we use

$$\hat{v}_i^\theta = \sum_{j=1}^{N_v} b_j B(y_i^\theta - y_j), \quad i = 1, \ldots N_u, \tag{5}$$

where B is a cubic B-spline and $\{b_j\}$ are the corresponding spline coefficients. To ensure exact interpolation, the B-spline coefficients are obtained by appropriately pre-filtering the original image $\{\tilde{v}_j\}$ using techniques described in [14]. Similarity metrics Ψ employing this model can be written as

$$\Psi(\theta) = \Psi(\{\tilde{u}_i, \hat{v}_i^\theta\}_{i=1}^{N_u}). \tag{6}$$

Assuming differentiability and using the chain rule, the gradient of Ψ is given by

$$g(\theta) \triangleq \nabla_\theta \Psi(\theta) = \sum_{i=1}^{N_u} \frac{\partial \Psi(\theta)}{\partial \hat{v}_i^\theta} \nabla_\theta \hat{v}_i^\theta. \tag{7}$$

where $\nabla_\theta = [\frac{\partial}{\partial \theta_1}, \frac{\partial}{\partial \theta_2} \ldots \frac{\partial}{\partial \theta_p}]$ denotes the gradient operator. To accelerate the gradient computation, a random subset of image voxels is typically drawn from a uniform sampling distribution [3], [7]. Thus any voxel pair is equally likely to be used to approximate the gradient, ensuring that the resulting approximation is unbiased.

Reducing the variance of this approximation (without introducing any bias) will not only improve the convergence of the SA optimizer but may also facilitate the use of smaller sample-sizes. This may be possible by using IS to encourage denser sampling from image regions that strongly influence the gradient given by (7). These 'important' image regions can be identified by differentiating (5):

$$\nabla_\theta \hat{v}_i^\theta = \left\{ \sum_{j=1}^{N_v} b_j \dot{B}(y_i^\theta - y_j) \right\} [\nabla_\theta y_i^\theta], \tag{8}$$

where $\dot{B}(y) = \nabla_y B(y)$, $y \in \mathbb{R}^3$ is the 1×3 vector gradient of the B-spline kernel. The term in the braces contains the directional gradients or edges of the homologous intensity image along the three coordinate axes. Recalling (7), only voxel intensities that lie on an edge in the homologous image $\{\hat{v}_i^\theta\}$ will contribute significantly to $g(\theta)$.

To see the importance of edges in the reference image we consider registration by swapping the two images, i.e., treating $\{\tilde{v}_j\}$ as the reference image and $\{\tilde{u}_i\}$ as the homologous image. This corresponds to finding an 'inverse' warp. In this case, the continuous function $u(.)$ will be modeled using an interpolation kernel. Repeating the above analysis, we see that edges in the swapped reference image $\{\hat{u}_j^\theta\}$ will now be vital in the gradient calculation. This suggests that our importance sampling scheme should follow a distribution that emphasizes edges in both the reference and the homologous images.

At the kth SA iteration with parameter guess $\theta = \theta_k$, we base the design of our θ-dependent sampling distribution P_s^θ on the edge magnitudes of the two intensity images. We choose the probability that a voxel pair with coordinates (x_i, y_i^θ) is selected as follows:

$$P_s^\theta(i) \triangleq \frac{e_i^\theta}{\sum_{j=1}^{N_u} e_j^\theta}, \quad i = 1, 2, \ldots N_u, \tag{9}$$

where

$$e_i^\theta \triangleq \begin{cases} \frac{s_i}{\sum_{j=1}^{N_u} s_j} + \frac{t_i^\theta}{\sum_{j=1}^{N_u} t_j^\theta}, & \text{if } \frac{s_i}{\sum_{j=1}^{N_u} s_j} + \frac{t_i^\theta}{\sum_{j=1}^{N_u} t_j^\theta} \geq T \\ \epsilon & \text{else.} \end{cases}$$

In the above equation $\{s_i\}_{i=1}^{N_u}$ and $\{t_i^\theta\}_{i=1}^{N_u}$ are approximate edge magnitudes of the reference and interpolated homologous images respectively. T is a user-defined edge threshold and $\epsilon \in (0, T]$.

The minimum probability that a voxel is used in the gradient approximation is given by the parameter ϵ. We choose ϵ to be a positive non-zero constant, so that in the limit of a large number of IS draws, all voxel-pairs will contribute to the SA optimization scheme. The threshold T may be tailored to remove spurious noise induced edges from the sampling distribution. If the normalized edge magnitudes in both images were all smaller than T, then the sampling distribution would become uniform with each voxel pair having an equal chance of being selected.

Let $(x_i, y_i^\theta); i \in S$ where $S \subset \{1, 2, \ldots N_u\}$, be coordinates of voxel pairs belonging to the small random subset S, chosen according to $P_s^\theta(i)$. Then the approximate gradient used in (2) is given by

$$\hat{g}(\theta_k) = \sum_{i \in S} \frac{1}{w(i)} \frac{\partial \Psi(\theta)}{\partial \hat{v}_i^\theta} \nabla_\theta \hat{v}_i^\theta \Big|_{\theta = \theta_k} \tag{10}$$

where $w(i) = N_u P_s^\theta(i)$ and $\theta = \theta_k$ is the warp parameter guess at the kth SA iteration. The voxel pairs in random subset S follow the non-uniform sampling distribution given by (9). Such non-uniform samples may yield a biased

gradient estimate. However, by using the weights $w(i)$ to appropriately weight each voxel pair, we can ensure that the resulting gradient approximation in (10) is unbiased. This approximate gradient uses only $|S| \ll N_u$ voxel pairs; hence the time consuming sum in (7) is evaluated only at these $|S|$ sample points.

Interestingly, Sabuncu et al. [15] recently developed an edge-dependent sampling scheme to reduce the approximation error in their Euclidean Minimum Spanning Trees (EMST) based registration. However, results were demonstrated only for rigid registration of 2D brain images. Further, they did not study the variance-bias properties of their approximation and assigned the same weight to all samples.

D. Optimization Scheme

As discussed previously both Step-SA and Samp-SA can be used to estimate the unknown warp parameters. Our previous empirical results [16] comparing registration of simulated brain data indicated that under identical conditions Samp-SA has faster initial convergence than Step-SA; however, Step-SA appeared to be more stable at later iterations. Two schemes combining the advantages of these SA methods resulted in faster nonrigid registration: (i) an 'Hybrid-SA' scheme that started with Samp-SA for a fixed number of iterations and then switched to Step-SA and (ii) a 'Pyramid-SA' scheme that employed a variable combination of step and sample-sizes using a multi-resolution pyramid approach. Because of the prevalence of pyramid optimization schemes and their empirically demonstrated robustness to local minima [2], [3], we used Pyramid-SA for all experiments in this paper.

In our experiments all levels of Pyramid-SA used cubic B-spline representations of both images. Lower levels of the pyramid used coarse image approximations with small amounts of data to obtain initial warp estimates. These warp estimates were then refined at higher levels of the pyramid using more precise image representations by including more intensity data. Since coarse image approximations are accompanied by a loss of detail, low level warp estimates capture gross global alignment and are explained using fewer parameters. As image detail increases with pyramid levels, the warps become more elaborate and depend on a larger number of parameters. Thus successive levels of the pyramid use an increasing number of intensity pairs to estimate the similarity metric. In an SA framework, this corresponds to implicitly increasing the sample-size between each level of the pyramid. 'Optimal' warp parameters within each pyramid level were estimated using Step-SA. For simplicity we call this optimization scheme 'Pyramid-SA'. In lieu of a gradient-dependant stopping criterion, we used a fixed number of SA iterations at each level of the pyramid. The exact number of Step-SA iterations at each level of our Pyramid-SA scheme was chosen heuristically.

When the number of unknown warp parameters is very small, it may be sufficient to empirically identify a single step-size value for Step-SA algorithms. However for large-dimensional vector valued parameters, the optimal step-size for each vector component may vary widely. To remedy this, we adopted an adaptive step-size estimation technique proposed in [17]. Let θ_k be the estimate of warp parameters at iteration k, with elements $\{\theta_k^i\}, i = 1, \ldots p$. The adaptive step-size strategy assumes that for a stationary point θ_* of the similarity measure, rapid changes in the sign of $(\theta_k^i - \theta_*^i) - (\theta_{k-1}^i - \theta_*^i) = \theta_k^i - \theta_{k-1}^i$ indicate that θ_k^i is closer to its optima. Similarly, fewer sign changes are indicative of a greater distance from θ_*^i. Thus the step-size associated with the ith warp parameter component is

kept inversely proportional to the number of sign changes of $\theta_k^i - \theta_{k-1}^i$. Our implementation estimates the step-size for the ith component θ_k^i as follows: $a_k^i = a_0/(A + Q_k^i)$, where Q_k^i is the number of sign changes in $\{\theta_m^i - \theta_{m-1}^i\}$, $m = 2, \ldots k$ and $Q_1^i = 0$. A and a_0 are positive non-zero constants. Such step-size sequences were shown to be convergent in [17]. While many choices of A and a_0 values are valid in theory, using a larger a_0 may boost SA performance by yielding larger step-sizes at later iterations [18]. However a larger a_0 may also result in instabilities at earlier iterations. It was observed in [18] that incorporating 'stability constant' $A \leq 0.1 \times (\text{Number of SA Iterations})$ could avoid such fluctuations in earlier SA iterations, allowing the use of larger a_0 values. For all experiments in Sec. III, we found that Pyramid-SA with two pyramid levels worked well, with $A = 10$ and less than 400 Step-SA iterations at each level of the pyramid.

E. Implementation Issues

For IS to be advantageous in an image registration application, it is crucial to design a meaningful sampling distribution that requires minimal computational effort. The sampling distribution P_s^θ depends on the changing warp parameter estimates through $\{t_i^\theta\}_{i=1}^{N_u}$, so it has to be recomputed with significant variations in the SA estimates of θ. Thus it is important to use a fast and simple approximation of the edge maps. Since the reference image does not change throughout the registration, we pre-compute its (fixed) edge map $\{s_i\}_{i=1}^{N_u}$. However the homologous image geometry changes with updates in θ and corresponding edge magnitude values need to be recomputed. For large homologous images, edge maps based on higher order kernels such as the cubic spline in (5) can be computationally expensive. Hence we approximate edge magnitudes using fast first order finite central differences of the intensity images along each image dimension.

The sampling distribution (9) gives equal importance to the normalized edge magnitude maps of both the reference and the homologous image. In the early stages of the registration scheme, the reference and homologous images may be strongly mis-aligned. Hence it is important to frequently update the homologous image's edge map during initial iterations, so as to accurately emphasize all the 'important' mis-aligned regions in both images. However, towards the final stages of the registration algorithm, we can expect both images to be better aligned. That is, many of the homologous image edges will now coincide with those of the reference image. Thus, it may be computationally advantageous to update the homologous image edge map sparingly at later iterations. Further, the coarse-to-fine framework of the Pyramid-SA scheme in Sec. II-D inherently results in coarse scale changes in the warp estimate at lower levels of the pyramid, while finer warp adjustments occur at higher pyramid levels. At each iteration, coarse scale warp changes are more likely to significantly affect the edge map than finer refinements. Hence, we update the sampling distribution frequently at lower Pyramid-SA levels and increase the number of iterations m between updates as the optimizer switches to higher levels. SA algorithms are characterized by small steps along random search directions. Thus the sampling distribution P_s^θ is updated every m iterations to reflect the average change in these m warp estimates. At pyramid level $l = 1, 2, \ldots$ we used $m = 2^l$.

Lastly, at every update, the approximate homologous image edge map need be recomputed only at locations where the effective deformation is large enough to significantly change the edge magnitude. That is, we incrementally

(a) Example Sampling Distribution (b) Importance Sampling (c) Uniform Sampling

Fig. 1. Comparison of samples obtained using the sampling distribution given by (9) versus samples obtained by Uniform sampling. Images were created when the algorithm was not near registration.

update our finite central difference based edge estimate only at geometric coordinates that move more than the dimensions of a voxel in any direction on average. These measures ensure that the overhead required to compute and update the sampling distribution is reasonably small. Further, this fractional overhead reduces steadily with increasing sample-sizes. Fig. 1 shows the sampling distribution and corresponding samples obtained using importance sampling for registration of simulated brain datasets.

III. Results

We demonstrate the use of IS for image registration using both simulated and real data. Results include pair-wise monomodality and multimodality registration using two common intensity-based similarity metrics. All registration results using IS-based Pyramid-SA (IS-SA) and US-based Pyramid-SA (US-SA) described here employed the optimization framework detailed in Sec. II. For comparison, registration was also performed using deterministic Gradient Descent (GD) in the same multi-resolution pyramid framework. GD used all image voxels to compute the analytical gradient at each iteration. All three methods utilized multi-resolution representations of both images using cubic splines and estimated deformable warps based on B-splines.

A. Behavior of IS-SA with Variations in Step-size

A limitation of SA approaches is their sensitivity to tuning parameters such as step-sizes. If the sampling distribution P_s^θ designed in (9) reduces the variance of $\hat{g}(\theta)$, IS-SA can be expected to have an increased tolerance to variations in step-sizes. Simulated datasets were used to compare the behavior of multi-modal registration using IS-SA and US-SA with various step-sizes.

Mutual Information (MI) based registration was performed between $180 \times 260 \times 60$ T1 and PD-weighted simulated MR volumes with $1 \times 1 \times 3$ mm^3 voxels, obtained from ICBM [19]. A plug-in MI estimate between the two images,

given by

$$\begin{aligned}\Psi_{\text{mi}}(\theta) = \quad & -\sum_{l=1}^{L} \hat{P}_u(g_l)\log(\hat{P}_u(g_l)) \\ & -\sum_{m=1}^{M} \hat{P}_v(h_m;\theta)\log(\hat{P}_v(h_m;\theta)) \\ & +\sum_{m=1}^{M}\sum_{l=1}^{L} \hat{P}_{uv}(g_l,h_m;\theta)\log(\hat{P}_{uv}(g_l,h_m;\theta)),\end{aligned} \tag{11}$$

was used as the similarity metric. $\hat{P}_v(h_m;\theta)$ is the approximate probability that a homologous intensity voxel $\hat{v}_i^{\theta} \in [h_m-\eta, h_m+\eta]$; $\hat{P}_u$ and $\hat{P}_{uv}$ are defined similarly over intensity levels $g_l = g_1, g_2, \ldots, g_L$ and $h_m = h_1, h_2, \ldots, h_M$. These sets of intensity levels $\{g_l\}_1^L$ and $\{h_m\}_1^M$ are chosen to span the dynamic intensity range of the reference and homologous images respectively. Our use of gradient-based optimizers requires that we approximate these pdfs using differentiable kernel density estimates [20].

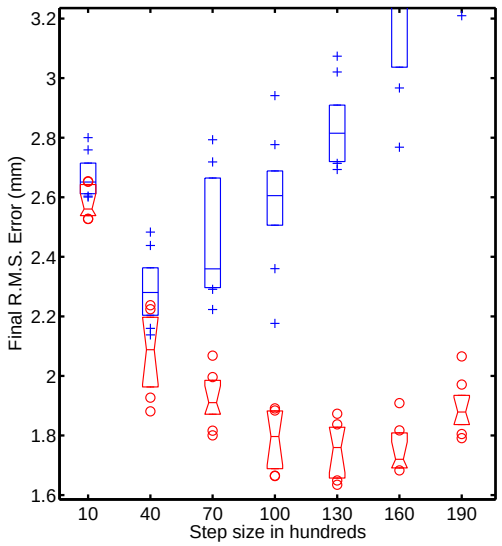
All results using IS-SA optimization schemes in this section used the sampling distribution given by (9). We applied a known synthetic warp $T(.)$ derived using radial blobs of varying severity to the T1 volume, yielding ground truth coordinates $T(x_i), i = 1, \ldots, N_u$. This warped volume was treated as the reference, while the unchanged PD volume was the homologous image. B-spline warps $T_{\hat{\theta}}(.)$ were estimated by mapping the homologous volume onto the reference volume. Independent realizations of Gaussian noise $N(0,9)$ were added to both images prior to the registration runs. Quality of the estimated warp $\{T_{\hat{\theta}}(x_i)\}_{i=1}^{N_u}$ was evaluated using the RMS error between the warp estimate and ground-truth:

$$\text{RMS error} = \sqrt{\frac{1}{N_u}\sum_{i=1}^{N_u} \|T(x_i) - T_{\hat{\theta}}(x_i)\|^2}. \tag{12}$$

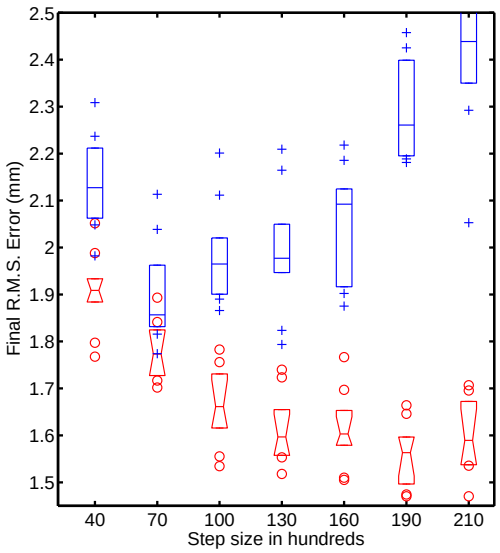
A two level Pyramid-SA scheme was used to register the two datasets. Level one used 64 histogram bins, a B-spline control point spacing of $16 \times 16 \times 8$ voxels and both images were down-sampled by a factor of two in all dimensions. The second level had 128 histogram bins, an $8 \times 8 \times 4$ voxels B-spline control point spacing and no down-sampling. Both levels implemented 150 and 250 iterations of Step-SA respectively and used only a fixed percentage of all available voxel pairs at that level.

The step-size a_k^i, corresponding to component θ_k^i of the warp parameters' estimate at iteration k, was $a_k^i = a_0/(10 + Q_k^i), i = 1, 2, \ldots, p$. Where, Q_k^i was the number of sign changes in $\{\theta_m^i - \theta_{m-1}^i\}, m = 2, \ldots, k$. The parameter a_0 in the step-size sequence remains to be chosen. To study the effect of varying step-size parameter a_0, warp estimates from 10 registration runs were obtained using IS and US, for systematically increasing values of a_0 from 1000 up to 25000 in increments of 3000. This process was repeated for four different sample sizes of $0.25, 0.5, 1$ and 2 percent respectively. Fig. 2 compares statistics of the final RMS errors obtained using the two sampling strategies for a fixed CPU time. As hypothesized, IS-SA yields lower errors than US-SA over the entire range of step-sizes.

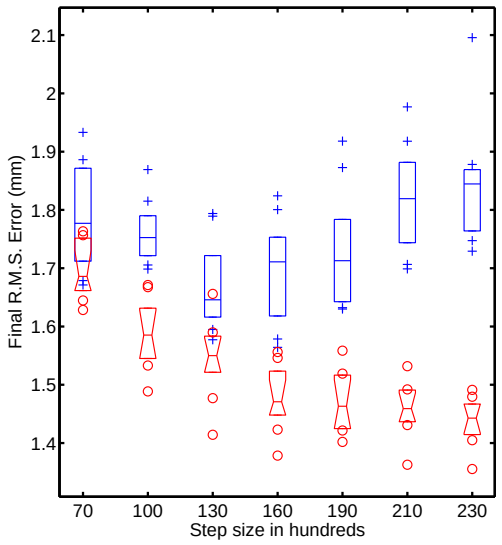
Empirically, IS-SA was significantly less sensitive to step-size variations, while consistently giving more accurate warp estimates. Further, US-SA required larger sample sizes to achieve accuracies comparable to those using IS.



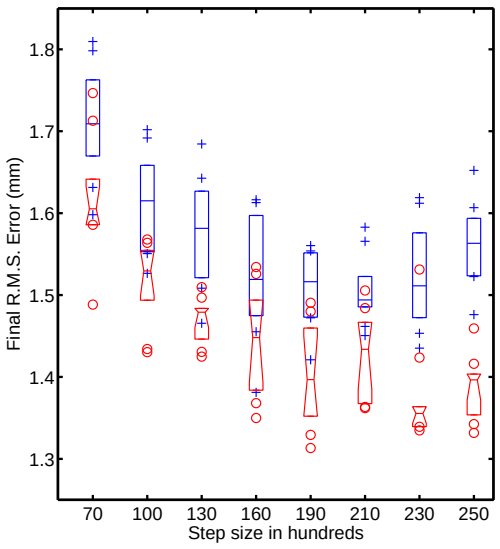
(a) Each pyramid level used 0.25% of all available voxels.



(b) Each pyramid level used 0.5% of all available voxels.



(c) Each pyramid level used 1% of all available voxels.



(d) Each pyramid level used 2% of all available voxels.

Fig. 2. Comparison of the performance of IS-SA (red/notched) versus US-SA (blue/plain) with variations in step-sizes. Figures show RMS error statistics for 10 nonrigid multimodality registration runs at seven step-sizes and four ($0.25, 0.5, 1$ and 2%) sample-sizes. The line at the center of each boxplot shows the median RMS error value and top and bottom edges are the 75 and 25 percent quantile RMS errors. 'Outliers' are shown by (o) for IS and by (+) for US. IS does significantly better than US at all four sample-sizes. Specifically, IS results in lower variance values and shows better tolerance to variations in step-sizes. Trends in the four plots indicate that the performance of both sampling strategies will become comparable with an increase in sample-size.

As sample-sizes increase both IS and US will capture similar levels of image complexity making their performance comparable. The minimum sample-size beyond which both sampling methods give similar results will depend on the complexity of the datasets. In general, US will be effective at smaller sample-sizes when image edge features are roughly uniformly dispersed.

B. Application to Human Data

Encouraged by the observations made in the previous section, we used IS to register human datasets. Intensity-based registration using B-spline warps was applied to align CT inhale and exhale lung datasets from 8 subjects. These CT scan pairs were obtained using a helical CT scanner (CT/I, General Electric, Milwaukee, WI) with $0.187\times0.187\times0.5$ cm^3 voxels. Each scan pair was acquired during coached voluntary breath-hold periods of 18 to 35 secs; the first scan at normal exhale followed by one at normal inhale. A more detailed description of the data can be found in [21].

Monomodality registration was performed using the negative of Sum of Squared Differences (SSD) as a similarity metric. In this case, both the reference and homologous images are assumed to be noisy realizations drawn from the same continuous function. Let the reference image be given by a set of noisy samples $\{\tilde{u}_i\}_{i=1}^{N_u}$. Then the negative SSD similarity metric is

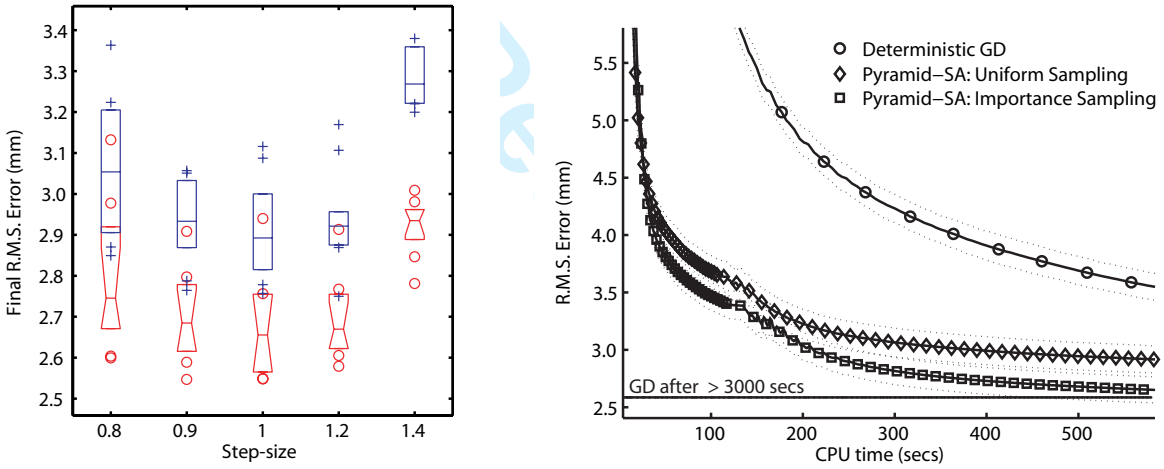
$$\Psi_{\text{SSD}}(\theta) = -\frac{1}{N_u}\sum_{i=1}^{N_u}\left(\tilde{u}_i - \hat{v}_i^{\theta}\right)^2, \tag{13}$$

where the interpolated homologous image $\{\hat{v}_i^{\theta}\}_{i=1}^{N_u}$ is given by (5). Differentiating the above expression shows that image edges are important to the gradient of Ψ_{SSD}. To ensure that Ψ_{SSD} was not affected by inherent differences in the scale of intensities of the two images, both images were normalized to have the same intensity ranges prior to registration.

Step-size Training

Effective use of US-SA or IS-SA to register a population of real datasets requires an efficient strategy to estimate the step-size parameter a_0. Here we outline a simple procedure to estimate this a_0 value using a single randomly chosen dataset from the target CT population. In the absence of known ground truth, B-spline warp estimates obtained using deterministic GD optimization were treated as the pseudo ground-truth. This is a reasonable assumption since the goal of our SA algorithms is to use only a small subset of strategically selected image voxels to attain registration accuracy comparable to that using GD with all image voxels. To mitigate local minima, registration estimates from multiple runs of a GD algorithm were used. Each run was initialized using a small randomly generated warp. The final registration estimate corresponding to the largest similarity metric value was treated as the best attainable warp. For a given sample-size, optimal a_0 values using both IS-SA and US-SA were chosen to consistently find warp estimates that yielded the smallest RMS error values with respect to this pseudo ground-truth warp.

For training purposes, we employed a two-level pyramid registration scheme. Level 1 downsampled the images by a factor of 2, estimated B-spline warps with a $16 \times 16 \times 8$ voxels control point spacing and used a_0 as the step-size parameter. The second level used no downsampling, a $8 \times 8 \times 4$ B-spline control point spacing and the step-size parameter was $1.5 \times a_0$. Each level used 1% of the total available voxels at that level. Ten wrap estimates were obtained using both IS-SA and US-SA for a set of five different a_0 values. Each registration run was terminated after 10 mins and at every iteration we recorded RMS errors of the estimated B-spline warp with respect to the pseudo ground-truth warp. Step-size parameter value $a_0 = 1$ was found to yield the best results for both SA methods. Fig. 3(a) shows statistics of RMS error values for all 10 IS-SA and US-SA registration runs at all five a_0 values. Fig. 3(b) shows speed and accuracy comparisons of GD, IS-SA and US-SA (both using $a_0 = 1$) with respect to the pseudo ground-truth warp. All subsequent SA based registrations were performed using this trained pyramid scheme with $a_0 = 1$.



(a) Step-size parameter (a_0) estimation.

(b) Speed and accuracy comparison of IS-SA, US-SA and GD.

Fig. 3. Comparison of the speed and accuracy of IS-SA (red/notched) and US-SA (blue/plain) for registration of CT Lung data. The optimal step-size parameter a_0 was empirically chosen to consistently produce warp estimates closest to the pseudo ground-truth warp in an RMSE sense. Fig. 3(a) shows that $a_0 = 1$ was the best value for both methods. The line at the center of each box-plot is the median RMS error, while top and bottom edges are 75 and 25 percent quantiles. Outliers are represented by ($\circ$) for IS-SA and ($+$) for US-SA. Fig. 3(b) shows how the speed and accuracy of the best IS-SA and US-SA schemes ($a_0 = 1$ and sample-size $= 1\%$) compare with those using GD (sample-size $= 100\%$) on average. Dotted lines are ± 1 standard deviation plots.

Validation

To gauge the performance of IS-SA and US-SA based on the trained pyramid scheme described above, we applied both methods to register all 8 CT inhale-exhale lung scan pairs. To quantify registration accuracy, six expert identified feature points were used per scan pair. These features included both bronchial and vascular bifurcations. For each subject, registration was performed by treating the exhale scan as the reference and the inhale scan as

the homologous dataset. Following registration, the estimated B-spline warp was used to transform the six exhale feature point coordinates to obtain predicted inhale feature point coordinates. The average of the Euclidean distance between the coordinates of each predicted and expert identified inhale feature point was used as an error metric to quantify registration accuracy for each dataset.

Since in reality we wish to replace a single GD registration run by a single SA registration run it is important that the method of choice give consistently good warp estimates with as little variance as possible. To empirically demonstrate the estimate variance associated with both SA methods, each CT dataset registration was repeated ten times. For comparison each dataset was also registered using GD. Each of the ten GD repetitions was initialized with a small random independently generated warp. Each SA registration run was completed in approximately 5 to 8 mins on a moderate PC running C++ code; in contrast, each successful GD registration required about 30 to 90 mins. Fig. 4 summarizes statistics of the resulting feature point error metric for all ten registration warp estimates using IS-SA and US-SA for all 8 datasets. In general IS-SA resulted in better accuracy than US-SA and showed a reduction in estimate variance.

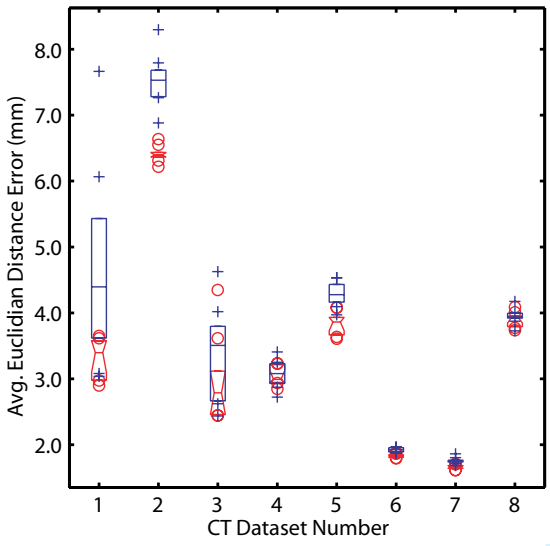


Fig. 4. Comparison of the accuracy and variation in trained IS-SA (red/notched) versus US-SA (blue/plain) registration using expert identified feature points for CT inhale-exhale lung data. The line at the center of each box-plot is the median error metric, while top and bottom edges are 75 and 25 percent quantiles. Outliers are represented by (○) for IS-SA and (+) for US-SA. Dataset 5 was used in the training step.

The average Euclidian distance between the expert identified exhale and inhale feature points can be used as some measure of the severity of the initial deformation. Table I indicates that for datasets with larger deformations (datasets 1, 2 and 3) IS-SA showed a marked improvement in accuracy over US-SA. For datasets with smaller deformations (datasets 6, 7 and 8) both methods performed comparably with IS-SA doing only slightly better than US-SA. The datasets are presented in order of decreasing initial deformation for ease of comparison. For most datasets IS-SA showed accuracy comparable to that using GD. Empirically, for datasets with larger deformations,

SA methods appeared to be less susceptible to local minima than GD. For datasets 1, 2 and 3 most repeated GD registration trials got stuck in local minima and terminated after 5 to 7 mins. These GD registrations resulted in poor inhale feature point predictions and were discarded as unsuccessful. In particular no GD registration run was successful for datasets 2 and 3, while only one run managed to escape local minima for dataset 1.

TABLE I

COMPARISON OF THE AVERAGE EUCLIDIAN DISTANCE ERROR FOR INHALE FEATURE POINTS PREDICTED USING US-SA, IS-SA AND GD.

Avg. Error	CT Dataset Number							
(mm)	1	2	3	4	5	6	7	8
Initial	15.10	14.52	13.31	11.73	9.13	8.62	7.77	6.89
Final								
US-SA	4.64	7.52	3.40	3.06	4.29	1.92	1.76	3.95
IS-SA	3.31	6.41	2.97	3.05	3.84	1.83	1.66	3.89
GD	3.14	-	-	2.15	3.29	1.95	2.12	3.63

IV. DISCUSSION AND CONCLUSION

We have developed and validated an importance sampling based stochastic approximation (IS-SA) approach to accelerate nonrigid image registration. We leveraged the significant influence of image edges on gradients of intensity-based similarity metrics to design an adaptive non-uniform sampling distribution that encourages sampling from these regions. Results for both synthetic simulations and real lung CT data show that registration using IS-SA can yield better speed and accuracy than SA schemes that use uniform sampling (i.e., US-SA). In particular, Fig. 2 shows that the number of samples required to attain a particular registration accuracy was halved by using IS-SA. For a fixed sample-size in Fig. 3(b) IS-SA was more than 2 times faster than US-SA on average. In contrast to approaches that replace or modify existing similarity metrics by explicitly incorporating image gradient-based terms [22], [23], our IS-based SA strategy can improve the speed and accuracy of a wider range of existing intensity-based registration methods without altering their similarity metrics (such as SSD, MI).

Correspondences between six expert identified bronchial and vascular bifurcations from each inhale-exhale CT scan pair were used in the validation procedure in Sec. III-B. While the selection of these bifurcations may have depended on edges, most of the voxels drawn in each IS-SA iteration using the sampling distribution (9) would not be near any bifurcation. Hence, we expect any bias toward IS-SA in the validation criterion to be small.

The use of SA methods in practical applications can be hindered by their dependence on the step-size parameter. To effectively apply these methods to populations of real data, we introduced a training strategy to empirically estimate a reasonable value for this step-size parameter in the absence of ground-truth. The training method uses only a single randomly chosen dataset from the target population and its corresponding ‘successful’ deterministic

GD registration warp estimate. This approach should be practical when several scans from the same protocol need to be registered. Finding automatic parameter selection methods for a single image pair is a challenging open problem.

Though we have demonstrated the efficacy of IS-SA only with B-spline warps, our framework is applicable to most other parametric non-rigid warp models. Specifically for more global warps (such as Thin-plate Splines) where each warp parameter depends on a larger number of image voxels, we expect to see more marked improvements in registration performance using IS-SA.

The data used here to demonstrate improvements in registration using IS-SA had few or sparse edges. As the percentage of edges increases it may be beneficial to use a more stringent criterion to retain fewer edges in the sampling distribution. More empirical experiments will be needed to quantify the approximate percentage of edges that need to be retained in such cases. In our implementation, the small random subset of samples S following the sampling distribution in (9) was drawn using the 'inverse pdf transform' sampling method. Alternatively, the samples in S may be drawn using a rejection sampling-like approach; especially when the datasets have a large percentage of edges. Further, an edge-based sampling strategy may not be the best choice for registration when one image has significant strongly demarcated structures absent from the other image(s).

The edge-based sampling distribution in (9) is not necessarily optimal. Since the gradient $g(\theta)$ in (7) depends on both $\nabla_\theta \hat{v}_i^\theta$ and $\frac{\partial \Psi(\theta)}{\partial \hat{v}_i^\theta}$; $i = 1, 2, \ldots N_u$, it may be possible to design alternative sampling distributions that emphasize image regions where both these terms are large. Finally, we note that a class of low discrepancy sequences, namely Highly Uniform Point-sets (HUPS), were used in [24] to improve the performance of uniform sampling based registration. A similar strategy, i.e., transforming such HUPS to obtain samples that follow the target sampling distribution in (9), may further augment the performance of importance sampling based registration.

ACKNOWLEDGMENT

The authors thank Michael Roberson for help with code and Marc L. Kessler and James Balter for access to the CT lung data.

REFERENCES

[1] C. R. Meyer, J. L. Boes, B. Kim, P. H. Bland, K. R. Zasadny, P. V. Kison, K. Koral, K. A. Frey, and R. L. Wahl, "Demonstration of accuracy and clinical versatility of mutual information for automatic multimodality image fusion using affine and thin plate spline warped geometric deformations," *Med. Im. Anal.*, vol. 1, no. 3, pp. 195–206, Apr. 1997.

[2] P. Thévenaz and M. Unser, "Optimization of mutual information for multiresolution image registration," *IEEE Trans. Im. Proc.*, vol. 9, no. 12, pp. 2083–99, Dec. 2000.

[3] D. Mattes, D. R. Haynor, H. Vesselle, T. K. Lewellen, and W. Eubank, "PET-CT image registration in the chest using free-form deformations," *IEEE Trans. Med. Imag.*, vol. 22, no. 1, pp. 120–8, Jan. 2003.

[4] D. Rueckert, P. Aljabar, R. A. Heckemann, J. V. Hajnal, and A. Hammers, "Diffeomorphic registration using B-splines," in *Medical Image Computing and Computer-Assisted Intervention*, 2006, pp. 702–9.

[5] G. Christensen, R. D. Rabbitt, and M. I. Miller, "Deformable templates using large deformation kinematics," *IEEE Trans. Im. Proc.*, vol. 5, no. 10, pp. 1435–47, Oct. 1996.

[6] J.-P. Thirion, "Image matching as a diffusion process: an analogy with Maxwell's demons," *Med. Im. Anal.*, vol. 2, no. 3, pp. 243–60, Sept. 1998.

[7] S. Klein, M. Staring, and J. P. W. Pluim, "Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines," *IEEE Transactions on Image Processing*, vol. 16, no. 12, pp. 2879–2890, Dec. 2007.

[8] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *Ann. Math. Stat.*, vol. 23, no. 3, pp. 462–6, Sept. 1952.

[9] J. C. Spall, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Trans. Auto. Control*, vol. 37, no. 3, pp. 332–41, Mar. 1992.

[10] H. Kushner and T. Gavin, "Extensions of Kesten's adaptive stochastic approximation method," *The Annals. of Statistics*, vol. 1, no. 5, pp. 851–61, Sept. 1973.

[11] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Stat.*, vol. 22, no. 3, pp. 400–7, Sept. 1951.

[12] P. Dupuis and R. Simha, "On sampling controlled stochastic approximation," *IEEE Trans. Auto. Control*, vol. 36, no. 8, pp. 915–24, Aug. 1991.

[13] S. Koonin, *Computational Physics*. California: Addison-Wesley, 1986.

[14] M. Unser, A. Aldroubi, and M. Eden, "B-spline signal processing: Part II—efficient design and applications," *IEEE Trans. Sig. Proc.*, vol. 41, no. 2, pp. 834–48, Feb. 1993.

[15] M. R. Sabuncu and P. J. Ramadge, "Gradient based nonuniform sampling for information theoretic alignment methods," in *Proc. Int'l. Conf. IEEE Engr. in Med. and Biol. Soc.*, vol. 3, 2004, pp. 1683–6.

[16] R. Bhagalia, J. A. Fessler, and B. Kim, "Gradient based image registration using importance sampling," in *Proc. IEEE Intl. Symp. Biomed. Imag.*, 2006, pp. 446–9.

[17] H. Kesten, "Accelerated stochastic approximation," *Ann. Math. Stat.*, vol. 29, no. 1, pp. 41–59, Mar. 1958.

[18] J. Spall, "Implementation of the simultaneous perturbation algorithm for stochastic optimization," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 34, no. 3, pp. 817–823, Jul 1998.

[19] D. L. Collins, A. P. Zijdenbos, V. Kollokian, J. G. Sled, N. J. Kabani, C. J. Holmes, and A. C. Evans, "Design and construction of a realistic digital brain phantom," *IEEE Trans. Med. Imag.*, vol. 17, no. 3, pp. 463–8, June 1998.

[20] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. New York: Wiley, 2001.

[21] M. M. Coselmon, J. M. Balter, D. L. McShan, and M. L. Kessler, "Mutual information based CT registration of the lung at exhale and inhale breathing states using thin-plate splines," *Med. Phys.*, vol. 31, no. 11, pp. 2942–8, Nov. 2004.

[22] E. Haber and J. Modersitzki, "Intensity gradient based registration and fusion of multi-modal images," in *MICCAI (2)*, 2006, pp. 726–733.

[23] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Image registration by maximization of combined mutual information and gradient information," *IEEE Trans. Med. Imag.*, vol. 19, no. 8, pp. 809–14, Aug. 2000.

[24] P. Thévenaz, M. Bierlaire, and M. Unser, "Halton sampling for image registration based on mutual information," *Sampling Theory in Signal and Image Processing*, vol. 7, no. 2, pp. 141–171, May 2008.

RESPONSE TO REVIEWERS' COMMENTS

Note: The page numbers referenced in our responses are actual manuscript page numbers (not 'Workflow' numbers).

Associate Editor: Pluim, Josien
Comments to the Author:
Please replace the in press reference [22] by the published one and conference papers by journal publications (if possible).

Reference [24] (previously, reference [22]) now cites the published work. Conference paper references [7] and [19] have been replaced by citations to corresponding journal publications.

Reviewer: 1
Comments to the Author
I am satisfied with the answers to the issues I was raising in my previous review, and also with the new version of the submission.

Reviewer: 2
Comments to the Author
The paper proposes a very practical algorithm that can potentially have a great impact on the implementation of similarity based image registration methods. This reviewer is satisfied with the changes the authors have made to clarify the paper based on the reviews. The notation is especially much clearer and easy to follow. Their expanded discussion includes further comparisons with other work and points to directions for interesting future research.

Reviewer: 3
Comments to the Author
I like to thank the authors for improving the quality of the paper. However, some issues (like the reflection of other methods, discussion of the ill-posedness, or proper stopping) are still unaddressed, see details below. Though I would like to see these points addressed, I consider them to be minor and don't want to further postpone publication. I assume the authors make a responsible decision on how to deal with those.

1) The reflection of image registration tools is very narrow and includes only parametric approaches. Variational as well as analytic approaches (like the Thin-plate Spline approach, where coefficients can be computed solving a least squares problem) are not reflected. Furthermore, it is stated that registration is an optimization problem, which out-rules flow approaches like Christensen's fluid registration or Thirion's demon approach.

R: "Lines 15-17 on page 1 were modified to stress that the paper focuses only on registration methods that estimate parameterized warps and that can be expressed as an optimization problem. Though we focus on parametric approaches, we believe that there are a significant number of image registration schemes based on parametric warps and optimization strategies that can benefit from our importance sampling approach."

O: The modification reads: "In this paper, we focus on image registration methods that estimate parameterized warp models" which does not reflect other approaches like the ones outlined in the 1st review. Please fix.

While image registration methods using both non-parametric and parametric warps are available, the optimization framework and importance sampling strategy outlined in our work was designed specifically for registration

algorithms that estimate parametrized warp models. Developing an IS strategy to accelerate non-parametric registration methods based on PDEs and optical flow constraints is out of the scope of this submission.

Lines 15-18 on page 1 now note that while non-parametric/flow-based registration algorithms are available, our work focuses only on parametric registration methods.

2) Proper references are missing at various places (like for example at p1 15: This ill-posed problem is ...). Ill-posedness is identified as a major issue in registration in the very beginning but not addressed subsequently. However, ill-posedness is particularly relevant to the topic, since the ratio between the number of used data and the number of used parameters can be interpreted as a measure for ill-posedness: using a sample size of one while looking for a thousand parameters may not be wise. A reflection of these issues is missing.

R: "To address the earlier comment, line 15 on page 1 was modified to restrict attention to registration schemes using parameterized warps. Of course the problem may still be ill-conditioned, but this is true of all image registration methods based on many parameters, and is not the main point of this work. We removed the distracting mention of ill-posedness. Further, the minimum probability that any voxel is chosen is given by $\epsilon > 0$ in Eq. 9. Thus, in the limit, all available voxel pairs (or data samples) will be used to approximate the gradient. That is, over time, all voxel pairs will contribute to the SA optimization scheme. Sec. II C, lines 43-46 on page 5, were revised to comment on this issue."

O: Ill-posedness is a key issue and reducing the information using subsampling might be critical. The point was to address and reflect this problem, not to remove the discussion. Please fix.

Though ill-posedness is a potential problem in most existing non-rigid registration methods, we believe that our importance sampling strategy does not further exacerbate this issue. We ensure that each image voxel has a non-zero probability of being used to approximate the gradient by choosing $\epsilon > 0$ in Eq. 9. Hence, over a large number of iterations, all voxel pairs will contribute to the gradient approximation. That is, over time, the loss in information due to subsampling can be considered to be small. We re-worded lines 43-46 on page 5 in Sec. II C to stress this point.

3) Fixed.

4) It is also well-known in the literature that gradient based optimization of course heavily relies on the gradient. Here, however, the gradient is replaced by a stochastic process (in addition to a finite difference approximation). Thus, one would at least expect a stopping criteria based on a vanishing expectation value of the gradient. This would require more computation time. None of these issues are discussed.

R: "The gradient of the similarity metric is approximated using Eq. 10 and computing the expectation of this approximation seems impractical. The number of Step-SA iterations at each level of our Pyramid-SA optimization scheme were chosen heuristically. Lines 45-48 on page 6 were added to Sec. II D to address this issue."

O: The lack of a proper stopping criterion, results in either being far from optimal or performing many possibly redundant steps, which is related to the time issue. Please reveal your heuristics.

We found that Pyramid-SA with two pyramid levels and less than 400 Step-SA iterations at each level of the pyramid worked well for all experiments described in Sec. III. This is noted in Sec. II D, lines 15-17 on page 7.

5) The weighting function in eq. (9) leads to another non-differentiability of the objective function. It is not explained nor discussed why to use this function, what are the benefits as compared to differentiable functions, and how to automatically identify the parameters.

R: "The gradient of the similarity metric $\nabla_\theta \Psi(\theta)$ is approximated by $\hat{g}(\theta)$ using Eq. 10. This corresponds to evaluating the analytical summation in Eq. 7 over a small random subset of image voxel pairs given by S. Please note that the actual similarity metric $\Psi(\theta)$ and its gradient $\nabla_\theta \Psi(\theta)$ are not altered in any way.

The sampling distribution P_s^θ Eq. 9 is used only to draw the random subset of voxel pairs S. Since these random samples follow a non-uniform distribution, the weights $w(.) = N_u P_s^\theta(.)$ are required to ensure that the resulting gradient approximation in Eq. 10 remains unbiased. Thus, these weights are a consequence of importance sampling and do not affect the differentiability of $\Psi(\theta)$. Lines 57-59 on page 5 and lines 4-7 on page 6 were added ...".

O: Sorry, I'm a little confused here. It is true that the equality in the continuous framework of (4) holds independent on the choice of w, but after discretization this may change, doesn't it?The basic idea is to replace the sum over all pixels by a sum over a smartly chosen subset S, right? Thus, your "gradient" g depend on S. It is actually nor clear to me, whether g is a true gradient of an objective function, which then of course also depends on S. Thus, changing the subset will change the objective in the optimization. Anyway, I'm fine with point.

To obtain the $k+1$th guess of the warp parameters $\theta = \theta_{k+1}$ using Eq. 2, we approximate the gradient $g(\theta) = \nabla_\theta \Psi(\theta)$ by $\hat{g}(\theta_k)$. This approximate gradient is computed at the current (kth) guess of the warp parameters $\theta = \theta_k$. Thus, the approximation $\hat{g}(\theta_k)$ is based on a new random subset S, drawn from the adaptive sampling distribution P_s^θ in Eq. 9 computed at $\theta = \theta_k$. That is, the random subset S used to compute $\hat{g}(\theta_k)$ depends on θ_k through $P_s^{\theta_k}$. However the properties of the actual similarity metric $\Psi(\theta)$ and its gradient $g(\theta)$ are not altered.

We modified line 27 on page 5 in Sec. II C to clarify that $\theta = \theta_k$ in Eq. 9. We also modified Eq. 10 to note that $\hat{g}(\theta_k)$ is computed at the current guess of warp parameters $\theta = \theta_k$.

The guideline for the parameters in (9) is still missing.

All the symbols and parameters in Eq. 9 are described in the lines immediately following Eq. 9.

In particular, line 39 on page 5, defines $\{s_i\}_{i=1}^{N_u}$ and $\{t_i^\theta\}_{i=1}^{N_u}$ as the approximate edge magnitudes of the reference and (interpolated) homologous images, respectively. These edge magnitudes are approximated using finite central differences (as noted in Sec. II E on page 7, lines 30-34) and are used to compute only the sampling distribution in Eq. 9. A discussion relating to the choice of the user-defined threshold T in presented in the paragraph following Eq. 9 in Sec. II C, lines 46-49 on page 5. Comments on the choice of ϵ appear at the beginning of the same paragraph.

3. Minor Objections ... are resolved.