Fast Nonparametric Mutual-Information-based Registration and Uncertainty Estimation

Mikael Agn[1(✉)] and Koen Van Leemput[1,2]

[1] Department of Health Technology, Technical University of Denmark, Lyngby, Denmark
miag@dtu.dk

[2] Martinos Center for Biomedical Imaging, MGH, Harvard Medical School, Boston, USA

Abstract. In this paper we propose a probabilistic model for multi-modal non-linear registration that directly incorporates the mutual information (MI) metric into a demons-like optimization scheme. In contrast to uni-modal registration, where the demons algorithm uses repeated spatial filtering to obtain very fast solutions, MI-based registration currently relies on general-purpose optimization schemes that are much slower. The central idea of this work is to reformulate an often-used histogram interpolation technique in MI implementations as an explicit spatial interpolation step within a generative model. By exploiting the specific structure of this model, we obtain a dedicated and fast expectation-maximization optimizer with demons-like properties. This also leads to an easy-to-implement Gibbs sampler to infer registration uncertainty in high-dimensional deformation models, involving very little additional code and no external tuning. Preliminary experiments on multi-modal brain MRI images show that the proposed optimizer can be both faster and more accurate than the free-form deformation method implemented in Elastix. We also demonstrate the sampler's ability to produce direct uncertainty estimates of MI-based registrations – to the best of our knowledge the first method in the literature to do so.

1 Introduction

An accurate and efficient way of non-linearly aligning two images with similar contrast properties is to minimize the sum-of-squared-differences (SSD) between them. The properties of the SSD criterion can be exploited to yield a dedicated optimization algorithm, the so-called demons algorithm [1,2], which repeatedly computes deformation "votes" at each voxel location, and spatially filters these votes to yield a spatially consistent deformation field. This results in fast optimizations of highly flexible, nonparametric deformation fields, taking only a few minutes on a standard desktop computer. Furthermore, the SSD criterion can

H. Greenspan et al. (Eds.): CLIP 2019/UNSURE 2019, LNCS 11840, pp. 42–51, 2019.
https://doi.org/10.1007/978-3-030-32689-0_5

be cast within a probabilistic modeling framework, making it possible to quantify registration uncertainty by approximating the relevant posterior probability distributions, using either variational [3–6] or sampling [7–10] methods.

In contrast to these methodological advances in deformable uni-modal registration, the *de facto* standard in the field of multi-modal registration using mutual information (MI) remains the free-form deformation approach [11], in which a parametric deformation model of B-spline basis functions is optimized with general-purpose optimization algorithms (e.g., [12,13]). This approach yields accurate registration results, but at a considerable computational cost when deformations with many degrees of freedom are needed (small spacing between the B-spline knots). Although attempts have been made to adapt faster, demons-like optimization schemes to the MI criterion [14–16], these methods have merely replaced the SSD-based demons "votes" with spatial MI gradients, a heuristic that does not necessarily optimize any specific objective function. Unlike the SSD criterion, MI does not currently have an associated probabilistic model [17], and consequently no principled way to quantify registration uncertainty.

In order to bring the SSD-specific techniques for uncertainty estimation and fast, nonparametric registration into the realm of MI-based registration, the contribution of this paper is threefold. First, we show that the partial volume interpolation technique for computing MI using fractional histogram counts [18,19] can be re-cast as a generative probabilistic model with an explicit spatial interpolation model. Second, we derive a tailor-made optimization algorithm that makes judicious use of latent variables in this model to obtain local "votes" of voxel-wise deformations that are subsequently regularized, allowing for a similar efficient optimization of nonparametric deformation models as in the demons algorithm. And third, using largely the same code base as the proposed optimizer, we also derive a practical technique for Monte Carlo sampling from the registration posterior, allowing for direct visualization and quantification of uncertainty in MI-based models. In contrast to existing methods for uncertainty estimation in uni-modal registration [3–10], this sampler does not involve variational approximations that may significantly underestimate uncertainty [9]; does not require tuning of various Metropolis-Hastings proposal distribution parameters; and can readily handle full 3D nonparametric deformation models with orders-of-magnitude more degrees of freedom than the sparse models used so far.

2 Generative Model

Let $\mathbf{u} = (u_1, \ldots, u_I)^{\mathrm{T}}$ denote an image with I voxels, where the intensities $u_i \in \{1, \ldots, L\}$ can take L discrete values. We model $\mathbf{u}$ as being generated from another image $\mathbf{v} = (v_1, \ldots, v_J)^{\mathrm{T}}$ with J voxels that we will refer to as "nodes", with intensities $v_j \in \{1, \ldots, K\}$ taken from K discrete levels, which we will call "classes". This is achieved by associating with each voxel i a spatial deformation d_i that maps that voxel to a spatial location $x_i + d_i$ in $\mathbf{v}$, where x_i denotes the voxel's initial position in $\mathbf{v}$. We also associate with each class k a class-specific

intensity distribution parameterized by $\boldsymbol{\theta}_k = (\theta_{k,1}, \ldots, \theta_{k,L})^{\mathrm{T}}$, where $\theta_{k,l}$ denotes the probability that the k^{th} class generates an intensity with value l. In the remainder, we assume periodic boundary conditions, and we only present the case in 1D, although the extension to higher dimensions is straightforward.

Using the notation $\mathbf{d} = (d_1, \ldots, d_I)^{\mathrm{T}}$ and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_k\}_{k=1}^K$ for the deformation field and the collection of all intensity distribution parameters, respectively, the generative process of $\mathbf{u}$ proceeds as follows: Let $\mathbf{n} = (n_1, \ldots, n_I)^{\mathrm{T}}$, $n_i \in \{1, \ldots, J\}$ denote latent node assignments, whereby each voxel i is associated with one node by centering a b^{th} order B-spline $\beta^b(\cdot)$ around its deformed position $x_i + d_i$, and using the B-spline value at each node location as the probability of that node being selected:

$$p(\mathbf{n}|\mathbf{d}) = \prod_{i=1}^{I} p(n_i|\mathbf{d}), \quad p(n_i = j|\mathbf{d}) = \beta^b\left(y_j - (x_i + d_i)\right),$$

where y_j denotes the spatial location of the j^{th} node. Subsequently, an intensity is drawn in each voxel from the distribution associated with the class of the selected node:

$$p(\mathbf{u}|\mathbf{n}, \boldsymbol{\theta}) = \prod_{i=1}^{I} p(u_i|n_i, \boldsymbol{\theta}), \quad p(u_i|n_i = j, \boldsymbol{\theta}) = \theta_{v_j, u_i}.$$

This induces a marginal distribution

$$p(\mathbf{u}|\mathbf{d}, \boldsymbol{\theta}) = \sum_{\mathbf{n}} p(\mathbf{u}|\mathbf{n}, \boldsymbol{\theta}) p(\mathbf{n}|\mathbf{d}) = \prod_{i=1}^{I} \left(\sum_{k=1}^{K} \pi_k(x_i + d_i) \theta_{k,u_i} \right)$$

where $\pi_k(z) = \sum_{j=1}^{J} [v_j = k] \beta^b(z - y_j)$ is a spatial map of the probability of class k, obtained as a B-spline expansion of the class assignments in the nodes of $\mathbf{v}$. Thus, the model effectively generates $\mathbf{u}$ by drawing, in each voxel i, a class from these probabilistic maps at location $x_i + d_i$, and subsequently generating an intensity from the selected class-specific intensity distribution.

The model is completed by specifying a prior encouraging spatial smoothness in the deformation field $p(\mathbf{d}) \propto \exp\left(-\frac{\gamma}{2}\|\boldsymbol{\Gamma}\mathbf{d}\|^2\right)$, where $\boldsymbol{\Gamma}$ is a $I \times I$ circulant matrix implementing a high-pass filter, and a prior $p(\boldsymbol{\theta}) = \prod_{k=1}^{K} \mathrm{Dir}\left(\boldsymbol{\theta}_k|\boldsymbol{\alpha}_0\right)$, where $\mathrm{Dir}\left(\cdot|\,\boldsymbol{\alpha}_0\right)$ denotes the Dirichlet distribution with parameters $\boldsymbol{\alpha}_0$. (A flat prior $p(\boldsymbol{\theta}) \propto 1$ is obtained by choosing $\boldsymbol{\alpha}_0 = \mathbf{1}$.)

3 Optimization

Registration of $\mathbf{u}$ with $\mathbf{v}$ can be obtained by fitting the model to the data: $(\widehat{\mathbf{d}}, \widehat{\boldsymbol{\theta}}) = \arg\max_{(\mathbf{d},\boldsymbol{\theta})} p(\mathbf{d}, \boldsymbol{\theta}|\mathbf{u})$ where $p(\mathbf{d}, \boldsymbol{\theta}|\mathbf{u}) \propto p(\mathbf{u}|\mathbf{d}, \boldsymbol{\theta}) p(\mathbf{d}) p(\boldsymbol{\theta})$. For this purpose, we propose an expectation-maximization (EM) algorithm that exploits the latent node assignments $\mathbf{n}$ in the model to achieve an efficient optimization strategy. In particular, we iteratively increase $\log p(\mathbf{d}, \boldsymbol{\theta}|\mathbf{u})$ from the current parameter estimates $(\tilde{\mathbf{d}}, \tilde{\boldsymbol{\theta}})$ by considering a lower bound $Q(\mathbf{d}, \boldsymbol{\theta}|\tilde{\mathbf{d}}, \tilde{\boldsymbol{\theta}}) \leq$

$\log p(\mathbf{d}, \boldsymbol{\theta}|\mathbf{u})$ that touches the objective function at the current estimates, i.e., $Q(\tilde{\mathbf{d}}, \tilde{\boldsymbol{\theta}}|\tilde{\mathbf{d}}, \tilde{\boldsymbol{\theta}}) = \log p(\tilde{\mathbf{d}}, \tilde{\boldsymbol{\theta}}|\mathbf{u})$, and subsequently optimizing this lower bound to find new parameter estimates:

$$(\tilde{\mathbf{d}}, \tilde{\boldsymbol{\theta}}) \leftarrow \arg\max_{(\mathbf{d}, \boldsymbol{\theta})} Q(\mathbf{d}, \boldsymbol{\theta}|\tilde{\mathbf{d}}, \tilde{\boldsymbol{\theta}}). \tag{1}$$

By design, this scheme guarantees that $\log p(\mathbf{d}, \boldsymbol{\theta}|\mathbf{u})$ is increased with every new iteration. The lower bound is constructed using Jensen's inequality, effectively "filling in" the unknown node assignments with their expectations:

$$\begin{aligned} Q(\mathbf{d}, \boldsymbol{\theta}|\tilde{\mathbf{d}}, \tilde{\boldsymbol{\theta}}) &\equiv \sum_{i=1}^{I}\sum_{j=1}^{J} w_{i,j}(\tilde{\mathbf{d}}, \tilde{\boldsymbol{\theta}}) \log\left[\frac{p(u_i|n_i = j, \boldsymbol{\theta})p(n_i = j|\mathbf{d})}{w_{i,j}(\tilde{\mathbf{d}}, \tilde{\boldsymbol{\theta}})}\right] + \log\left[\frac{p(\boldsymbol{\theta})p(\mathbf{d})}{p(\mathbf{u})}\right] \\ &\leq \sum_{i=1}^{I} \log \underbrace{\left[\sum_{j=1}^{J} \frac{p(u_i|n_i = j, \boldsymbol{\theta})p(n_i = j|\mathbf{d})}{w_{i,j}(\tilde{\mathbf{d}}, \tilde{\boldsymbol{\theta}})} w_{i,j}(\tilde{\mathbf{d}}, \tilde{\boldsymbol{\theta}})\right]}_{p(u_i|\mathbf{d}, \boldsymbol{\theta})} + \log\left[\frac{p(\boldsymbol{\theta})p(\mathbf{d})}{p(\mathbf{u})}\right] \\ &= \log p(\mathbf{d}, \boldsymbol{\theta}|\mathbf{u}), \end{aligned}$$

where

$$w_{i,j}(\mathbf{d}, \boldsymbol{\theta}) = p(n_i = j|u_i, \mathbf{d}, \boldsymbol{\theta}) = \frac{\theta_{v_j, u_i} \beta^b\left(y_j - (x_i + d_i)\right)}{\sum_{j'=1}^{J} \theta_{v_{j'}, u_i} \beta^b\left(y_{j'} - (x_i + d_i)\right)} \tag{2}$$

weighs the association of each voxel i with each of the j nodes, so that $\sum_{j=1}^{J} w_{i,j} = 1, \forall i$. Note that most $w_{i,j} = 0$, due to the limited spatial support of B-splines.

Finding new parameter estimates by solving Eq. (1) readily yields the following closed-form update for $\boldsymbol{\theta}$:

$$\tilde{\theta}_{k,l} \leftarrow \frac{N_{k,l} + (\alpha_0^l - 1)}{\sum_{l'=1}^{L}\left(N_{k,l'} + (\alpha_0^{l'} - 1)\right)} \quad \forall k, l, \tag{3}$$

where

$$N_{k,l} = \sum_{i=1}^{I}\sum_{j=1}^{J} [u_i = l][v_j = k] w_{i,j} \tag{4}$$

can be interpreted as the effective number of voxels with intensity l that were assigned to nodes of class k. The corresponding update for $\mathbf{d}$ is not given in closed form, but an efficient and accurate approximation can be obtained by observing that B-splines rapidly become more Gaussian-like as the order b increases: $\beta^b(z) \simeq \mathcal{N}(z|0, \sigma_b^2)$ for an appropriate choice of variance σ_b^2. Plugging in this approximation yields an objective function that is quadratic in $\mathbf{d}$, and that therefore has a closed-form solution:

$$\tilde{\mathbf{d}} \simeq \arg\min_{\mathbf{d}} \left[\sum_{i=1}^{I}\sum_{j=1}^{J} w_{i,j} \frac{(y_j - x_i - d_i)^2}{\sigma_b^2} + \gamma \mathbf{d}^{\mathrm{T}} \boldsymbol{\Gamma}^{\mathrm{T}} \boldsymbol{\Gamma} \mathbf{d}\right] = \mathbf{S}\boldsymbol{\delta}, \tag{5}$$

where

$$\mathbf{S} = \left(\mathbf{I}_I + \gamma\sigma_b^2\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Gamma}\right)^{-1}, \quad \boldsymbol{\delta} = (\delta_1, \ldots, \delta_I)^{\mathrm{T}}, \quad \delta_i = \sum_{j=1}^{J} w_{i,j}\, y_j - x_i. \tag{6}$$

Thus, in each voxel i a local "vote" for a displacement δ_i is made that would recover the distance between the node(s) the voxel associates with, and its actual position. These local votes are then spatially smoothed by a $I \times I$ matrix $\mathbf{S}$ that implements a shift-invariant low-pass filter, to give the new estimate for $\mathbf{d}$. In summary, the proposed EM optimizer iteratively cycles between updating the expected node assignments $w_{i,j}$ (Eq. (2)) and the estimates of $\boldsymbol{\theta}$ (Eq. (3)) and $\mathbf{d}$ (Eq. (5)). In the Appendix, we show that this optimization scheme effectively performs MI-based registration with partial volume interpolation [18,19].

In our implementation, we initialize the algorithm by setting $\theta_{k,l}^0 = 1/L, \forall k, l$ and $d_i^0 = 0, \forall i$, and we use cubic B-splines ($b = 3$) in order to obtain an accurate Gaussian approximation, where σ_b^2 is set so that $\mathcal{N}(0|0, \sigma_b^2) = \beta^b(0)$. For $\boldsymbol{\Gamma}$, we use a filter that computes local curvature using finite differences (a so-called bending energy or biharmonic model), and we use $\boldsymbol{\alpha}_0 = 2{\cdot}\mathbf{1}$. Since the smoothing matrix $\mathbf{S}$ is circulant, the filtering can be performed as element-wise multiplication in the Fourier domain. Implemented in ITK 5.0 and MATLAB 9.6 on an Intel Core i7-5930K computer with Intel MKL's FFTW library, one iteration of the EM algorithm takes around 3.5 s for images of size $256 \times 176 \times 256$.

4 Sampling

Rather than simply obtaining point estimates $(\widehat{\mathbf{d}}, \widehat{\boldsymbol{\theta}})$, the uncertainty of these estimates can be quantified by Monte Carlo sampling from the posterior distribution $p(\mathbf{d}, \boldsymbol{\theta}|\mathbf{u})$. Since $p(\mathbf{d}, \boldsymbol{\theta}|\mathbf{u}) = \sum_{\mathbf{n}} p(\mathbf{d}, \boldsymbol{\theta}, \mathbf{n}|\mathbf{u})$, we can again exploit the latent node assignments $\mathbf{n}$ to obtain an efficient sampling strategy: Starting from an initialization $(\mathbf{d}^{(0)}, \boldsymbol{\theta}^{(0)}) = (\widehat{\mathbf{d}}, \widehat{\boldsymbol{\theta}})$, a Gibbs sampler of $p(\mathbf{d}, \boldsymbol{\theta}, \mathbf{n}|\mathbf{u})$ is obtained by the iterative scheme

$$\begin{aligned}
\mathbf{n}^{(\tau+1)} &\sim p(\mathbf{n}|\mathbf{d}^{(\tau)}, \boldsymbol{\theta}^{(\tau)}, \mathbf{u}) = \prod_{i=1}^{I}\prod_{j=1}^{J} \left\{w_{i,j}(\mathbf{d}^{(\tau)}, \boldsymbol{\theta}^{(\tau)})\right\}^{[n_i=j]} \\
\boldsymbol{\theta}^{(\tau+1)} &\sim p(\boldsymbol{\theta}|\mathbf{u}, \mathbf{n}^{(\tau+1)}) = \prod_{k=1}^{K} \mathrm{Dir}\left(\boldsymbol{\theta}_k|\boldsymbol{\alpha}_k\right), \ \boldsymbol{\alpha}_k = (N_{k,1}^{(\tau+1)}, \ldots, N_{k,L}^{(\tau+1)})^{\mathrm{T}} + \boldsymbol{\alpha}_0 \\
\mathbf{d}^{(\tau+1)} &\sim p(\mathbf{d}|\mathbf{u}, \mathbf{n}^{(\tau+1)}) = \mathcal{N}\Big(\mathbf{d}\,|\,\mathbf{S}\boldsymbol{\delta}^{(\tau+1)}, \sigma_b^2\mathbf{S}\Big),
\end{aligned}$$

where $N_{k,l}^{(\tau+1)}$, and $\boldsymbol{\delta}^{(\tau+1)}$ are as defined in Eqs. (4) and (6) but with hard node assignments $w_{i,j} = [n_i^{(\tau+1)} = j]$. After discarding the first T_0 burn-in sweeps, the set $\{\mathbf{d}^{(\tau)}, \boldsymbol{\theta}^{(\tau)}\}_{\tau=T_0+1}^{T}$ contains $(T - T_0)$ valid samples of the target distribution

$p(\mathbf{d}, \boldsymbol{\theta}|\mathbf{u})$. Since the required computations are very similar to those of the EM algorithm ($\mathbf{d}^{(\tau+1)}$ can again be computed via the Fourier domain), implementing the sampler requires very little additional code, and the computation time of a single sweep is comparable to that of one EM iteration.

As in other work [3–9], the deformation regularization parameter γ can also be inferred automatically, rather than set by the user. When a non-informative gamma distribution $\mathrm{Gam}(\gamma|\alpha_0, \beta_0)$ with shape $\alpha_0 = 1$ and rate $\beta_0 = 0$ is used as a conjugate prior for γ, this can be accomplished by simply including a fourth step in the sampler: $\gamma^{(\tau+1)} \sim p(\gamma|\mathbf{d}^{(\tau+1)}) = \mathrm{Gam}(\frac{I}{2} + 1, \frac{1}{2}\|\boldsymbol{\Gamma}\mathbf{d}^{(\tau+1)}\|^2)$.

5 Experiments

In order to perform an initial, preliminary comparison of the performance of the proposed optimizer with that of the well-known free-form deformation method Elastix (v. 4.8) [13], we co-registered the T2-weighted brain scans of 6 healthy subjects to the T1-weighted scan of 10 other healthy subjects in the OASIS database [20]. We first segmented and bias field corrected each image (including an additional T1w scan for the 6 subjects with T2w) with a whole-brain segmentation tool [21], and affinely pre-registered each of the registration pairs with Elastix. Registration accuracy was quantified by computing the Dice scores between the T1-based segmentations for each of the 60 T2-T1 registration pairs, averaged over the 10 largest brain structures.

For Elastix, we varied the B-spline grid spacing between 4, 3.5, 3, and 2.5 voxels, and the number of iterations per multi-resolution level between 500 (which is the default) and 2000 (which is recommended for best results). For each parameter variation, we used recommended and default settings, with a 4-level multi-resolution strategy and 5000 off-grid samples per iteration. The proposed method used the same multi-resolution regime, and varied the deformation regularization parameter γ between 14.9, 8.7, 4.8 and 2.3 to achieve the same effective number of degrees of freedom (measured as the trace of the smoothing matrix $\mathbf{S}$ [22]) as the corresponding B-spline deformation models in Elastix. The number of iterations per resolution level was also varied, between 25, 50 and 100.

The middle row of Fig. 1 shows an example registration obtained with the proposed method when the maximum degrees of freedom and 100 iterations are used. The top row shows quantitative results for the proposed method across the various settings, along with the corresponding results obtained with Elastix. It can be seen that the computational burden (left plot) of the proposed method is independent of the flexibility of the deformation models, whereas for Elastix the computation time increases sharply as more degrees of freedom are added. Furthermore, whereas it is possible to obtain better Dice scores for both algorithms by increasing the number of iterations and the degrees of freedom (middle plot), the proposed method does so more effectively, reaching higher average accuracy levels in 2.5 min (25 iterations at the highest flexibility), than the best achievable performance of Elastix, taking around 23 min (2000 iterations at B-spline spacing of 3 voxels). Finally, we also show the percentage of voxels with a Jacobian

determinant lower or equal to zero for both methods (right plot), indicating that the proposed method's deformation model is better behaved.

The bottom row and the last image on the middle row of Fig. 1 illustrate the proposed sampler (initialized by the optimizer with $\gamma = 2.3$) across 4000 samples after a burn-in of 1000 sweeps, where γ was kept constant for the first 100 sweeps. The uncertainty (last image, middle row) is shown as the standard deviation (measured in voxels) in each of the three spatial directions, and is encoded as red for superior-inferior, blue for left-right, and green for anterior-posterior. The bottom row shows the estimated posterior distribution of γ, and two deformation field samples (also including superior-inferior), zoomed-in on a region of interest and color-coded according to displacement magnitude.

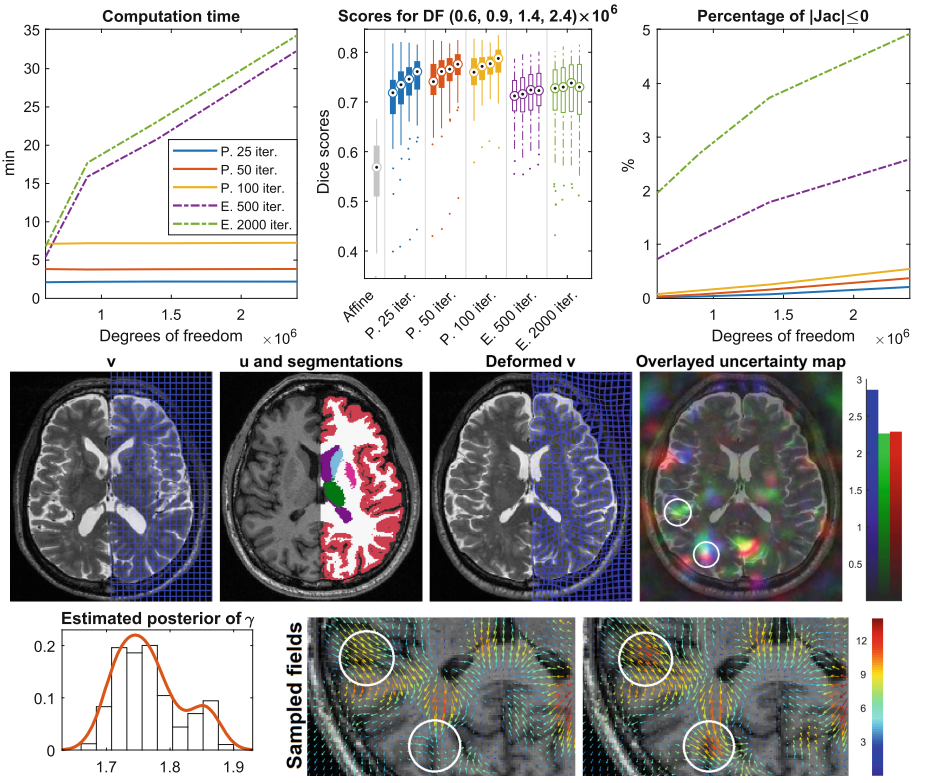

Fig. 1. Top: computation time; Dice scores; % of voxels with |Jacobian| ≤ 0. Middle: $\mathbf{v}$; $\mathbf{u}$ with segmentations partially overlayed; deformed $\mathbf{v}$; uncertainty map. Bottom: posterior of γ; two deformation field samples. P = Proposed method, E = Elastix.

6 Discussion

In this paper we have proposed a probabilistic model that directly incorporates the MI metric into a demons-like optimization scheme. We have shown that the resulting algorithm can potentially be more accurate and significantly faster than the free-form deformation method implemented in Elastix. We have also demonstrated that a Monte Carlo sampler, using largely the same code base, can directly produce uncertainty estimates in MI-based registration – to the best of our knowledge the first method in the literature to do so.

We note that although the generative model encodes MI in this paper, it can be used for a wide range of predictive distributions, including the Gaussian noise assumption underlying the SSD criterion. Although not reported here, preliminary experiments indicate that the proposed optimizer achieves comparable registration accuracies to the original demons algorithm [2] in this setting. The proposed sampler directly endows the demons algorithm with the first method to assess the uncertainty in its nonparametric deformation fields, the effective number of degrees of freedom of which is in the millions (compared to mere thousands in existing work for uncertainty estimation in registration [3–10]).

Given the close similarity between the two methods, in future work we plan to investigate whether the same update modification that makes the demons algorithm diffeomorphic [2] can also be used with the proposed optimizer.

Acknowledgments. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skło-dowska-Curie grant agreement No 765148; the Danish Council for Independent Research under grant number DFF611100291; and the NIH National Institute on Aging under grant number R21AG050122.

Appendix: Connection with MI-based registration

MI-based registration with partial volume interpolation can be interpreted as implicitly using the proposed generative model but with a different optimization strategy, in which EM is used to estimate $\boldsymbol{\theta}$ but not $\mathbf{d}$. In particular, $\widehat{\mathbf{d}}$ can also be estimated by optimizing $\log p(\mathbf{d}, \widehat{\boldsymbol{\theta}}_d|\mathbf{u})$ for $\mathbf{d}$ with a general-purpose optimizer, where $\widehat{\boldsymbol{\theta}}_d = \arg\max_{\boldsymbol{\theta}} \log p(\mathbf{d}, \boldsymbol{\theta}|\mathbf{u})$ involves an inner optimization that for each $\mathbf{d}$ estimates a matched $\widehat{\boldsymbol{\theta}}_d$ *de novo* from starting values $\theta^0_{k,l} = 1/L, \forall k, l$ by interleaving the EM Eqs. (2) and (3). When a flat prior $p(\boldsymbol{\theta}) \propto 1$ is used, the resulting effective registration criterion is then directly related to MI as follows:

$$\log p(\mathbf{d}, \widehat{\boldsymbol{\theta}}_d|\mathbf{u}) \simeq I\mathrm{MI}(\mathbf{d}) + \log p(\mathbf{d}) + \mathrm{const}, \tag{7}$$

where

$$\mathrm{MI}(\mathbf{d}) = \sum_{k=1}^{K}\sum_{l=1}^{L} n_{k,l} \log \frac{n_{k,l}}{n_k\, n_l} \quad \text{with} \quad n_{k,l} = \frac{N_{k,l}}{I},\ n_k = \sum_l n_{k,l},\ n_l = \sum_k n_{k,l}$$

is the MI criterion using partial volume interpolation [18,19], in which joint histogram counts $N_{k,l}$ are computed from fractional weights $\bar{w}^d_{i,j} = \beta^b\left(y_j - (x_i + d_i)\right)$ as in Eq. (4). To see why Eq. (7) holds, we can also write $\mathrm{MI}(\mathbf{d})$ as

$$\mathrm{MI}(\mathbf{d}) = \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J} \bar{w}^d_{i,j} \log \bar{\theta}^d_{v_j,u_i} - \underbrace{\sum_{l=1}^{L} n_l \log n_l}_{\text{const}} \quad \text{with} \quad \bar{\theta}^d_{kl} = n_{k,l}/n_k, \tag{8}$$

and, since $\log p(\mathbf{d}, \widehat{\boldsymbol{\theta}}_d) = Q(\mathbf{d}, \widehat{\boldsymbol{\theta}}_d | \mathbf{d}, \widehat{\boldsymbol{\theta}}_d)$,

$$\begin{aligned}\log p(\mathbf{d}, \widehat{\boldsymbol{\theta}}_d|\mathbf{u}) - \log p(\mathbf{d}) = &\sum_{i=1}^{I}\sum_{j=1}^{J} \widehat{w}^d_{i,j} \log \widehat{\theta}^d_{v_j,u_i} \\ &- \sum_i D_{KL}\left[p(n_i|u_i, \mathbf{d}, \widehat{\boldsymbol{\theta}}_d) \,\|\, p(n_i|\mathbf{d})\right] + \text{const},\end{aligned} \tag{9}$$

where $\widehat{w}^d_{i,j} = w_{i,j}(\mathbf{d}, \widehat{\boldsymbol{\theta}}_d)$ and $D_{KL}(.\|.)$ denotes the Kullback-Leibler (KL) divergence. Comparing Eqs. (8) and (9), and noting that $\bar{w}^d_{i,j}$ and $\bar{\boldsymbol{\theta}}_d$ are precisely the weights and estimate of $\boldsymbol{\theta}$ in the first iteration of the inner EM optimization, MI-based registration can therefore be interpreted as making a "lazy" attempt at measuring $\log p(\mathbf{d}, \widehat{\boldsymbol{\theta}}_d)$, using only a single iteration in the inner optimization of $\widehat{\boldsymbol{\theta}}_d$, and ignoring the KL divergence between the prior and the posterior node assignment distributions. In the special case where $p(n_i|\mathbf{d})$ takes only binary values $\{0, 1\}$, the approximation in Eq. (7) will be exact since the EM algorithm then immediately finds $\widehat{\boldsymbol{\theta}}_d$ in its first iteration and the KL divergence term vanishes. This will happen when B-splines of order $b = 0$ are used, or for first-order B-splines ($b = 1$) whenever the image grids of $\mathbf{u}$ and $\mathbf{v}$ perfectly align.

References

1. Thirion, J.P.: Image matching as a diffusion process: an analogy with Maxwell's demons. Med. Image Anal. **2**(3), 243–260 (1998)
2. Vercauteren, T., Pennec, X., Perchant, A., Ayache, N.: Diffeomorphic demons: efficient non-parametric image registration. NeuroImage **45**(1), S61–S72 (2009)
3. Simpson, I.J., Schnabel, J.A., Groves, A.R., Andersson, J.L., Woolrich, M.W.: Probabilistic inference of regularisation in non-rigid registration. NeuroImage **59**(3), 2438–2451 (2012)
4. Simpson, I.J.A., et al.: A bayesian approach for spatially adaptive regularisation in non-rigid registration. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013. LNCS, vol. 8150, pp. 10–18. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40763-5_2
5. Simpson, I.J., et al.: Probabilistic non-linear registration with spatially adaptive regularisation. Med. Image Anal. **26**(1), 203–216 (2015)
6. Le Folgoc, L., Delingette, H., Criminisi, A., Ayache, N.: Sparse bayesian registration of medical images for self-tuning of parameters and spatially adaptive parametrization of displacements. Med. Image Anal. **36**, 79–97 (2017)

7. Risholm, P., Samset, E., Wells, W.: Bayesian estimation of deformation and elastic parameters in non-rigid registration. In: Fischer, B., Dawant, B.M., Lorenz, C. (eds.) WBIR 2010. LNCS, vol. 6204, pp. 104–115. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-14366-3_10
8. Risholm, P., Janoos, F., Norton, I., Golby, A.J., Wells III, W.M.: Bayesian characterization of uncertainty in intra-subject non-rigid registration. Med. Image Anal. **17**(5), 538–555 (2013)
9. Le Folgoc, L., Delingette, H., Criminisi, A., Ayache, N.: Quantifying registration uncertainty with sparse bayesian modelling. IEEE Trans. Med. Imaging **36**(2), 607–617 (2016)
10. Pursley, J., et al.: A Bayesian nonrigid registration method to enhance intraoperative target definition in image-guided prostate procedures through uncertainty characterization. Med. Phys. **39**(11), 6858–6867 (2012)
11. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L., Leach, M.O., Hawkes, D.J.: Non-rigid registration using free-form deformations: application to breast MR images. IEEE Trans. Med. Imaging **18**(8), 712–721 (1999)
12. Modat, M., et al.: Fast free-form deformation using graphics processing units. Comput. Methods Prog. Biomed. **98**(3), 278–284 (2010)
13. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.: Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging **29**(1), 196–205 (2010)
14. Modat, M., Vercauteren, T., Ridgway, G.R., Hawkes, D.J., Fox, N.C., Ourselin, S.: Diffeomorphic demons using normalized mutual information, evaluation on multimodal brain MR images. In: SPIE Medical Imaging 2010: Image Processing, vol. 7623, p. 76232K (2010)
15. Lu, H., et al.: Multi-modal diffeomorphic demons registration based on pointwise mutual information. In: 2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 372–375 (2010)
16. Risser, L., Heinrich, M.P., Rueckert, D., Schnabel, J.A.: Multi-modal diffeomorphic registration using mutual information: application to the registration of CT and MR pulmonary images. In: Proceedings MICCAI Workshop PIA (2011)
17. Janoos, F., Risholm, P., Wells, W.: Bayesian characterization of uncertainty in multi-modal image registration. In: Dawant, B.M., Christensen, G.E., Fitzpatrick, J.M., Rueckert, D. (eds.) WBIR 2012. LNCS, vol. 7359, pp. 50–59. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-31340-0_6
18. Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P.: Multimodality image registration by maximization of mutual information. IEEE Trans. Med. Imaging **16**(2), 187–198 (1997)
19. Chen, H.M., Varshney, P.K.: Mutual information-based CT-MR brain image registration using generalized partial volume joint histogram estimation. IEEE Trans. Med. Imaging **22**(9), 1111–1119 (2003)
20. Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci. **19**(9), 1498–1507 (2007)
21. Puonti, O., Iglesias, J.E., Van Leemput, K.: Fast and sequence-adaptive whole-brain segmentation using parametric bayesian modeling. NeuroImage **143**, 235–249 (2016)
22. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer Series in Statistics, 2nd edn. Springer, New York (2009). https://doi.org/10.1007/978-0-387-84858-7