October 2007

Multi-start Method with Prior Learning for Image Registration

Gang Song
University of Pennsylvania, songgang@seas.upenn.edu

Brian B. Avants
University of Pennsylvania, avants@grasp.cis.upenn.edu

Jim C. Gee
University of Pennsylvania, gee@mail.med.upenn.edu

Recommended Citation

Gang Song, Brian B. Avants, and Jim C. Gee, "Multi-start Method with Prior Learning for Image Registration", . October 2007.

Multi-start Method with Prior Learning for Image Registration

Abstract

We propose an efficient image registration strategy that is based on learned prior distributions of transformation parameters. These priors are used to constrain a finite- time multi-start optimization method. Motivation for this approach comes from the fact that standard affine brain image registration methods, especially those based on gradient descent optimization alone, are affected by the initial search position. While global optimization methods can resolve this problem, they are are often very time consuming. Our goal is to build an explicit prior model of the gap between a typical registration solution and the solution gained by a global optimization method. We use this learned prior model to restrict randomized search in the relevant parameter space surrounding the initial solution. Global optimization in this restricted parameter space provides, in finite time, results that are superior to both gradient descent and the general multi-start strategy. The performance of our method is illustrated on a data set of 67 elderly and neurodegenerative brains. Our novel learning strategy and the associated registration method are shown to outperform other approaches. Theoretical, synthetic and real-world examples illustrate this improvement.

Comments

Multi-start Method with Prior Learning for Image Registration

Gang Song, Brian B. Avants and James C. Gee
Penn Image Computing and Science Laboratory
University of Pennsylvania
http://www.picsl.upenn.edu

Abstract

We propose an efficient image registration strategy that is based on learned prior distributions of transformation parameters. These priors are used to constrain a finite-time multi-start optimization method. Motivation for this approach comes from the fact that standard affine brain image registration methods, especially those based on gradient descent optimization alone, are affected by the initial search position. While global optimization methods can resolve this problem, they are are often very time consuming. Our goal is to build an explicit prior model of the gap between a typical registration solution and the solution gained by a global optimization method. We use this learned prior model to restrict randomized search in the relevant parameter space surrounding the initial solution. Global optimization in this restricted parameter space provides, in finite time, results that are superior to both gradient descent and the general multi-start strategy. The performance of our method is illustrated on a data set of 67 elderly and neurodegenerative brains. Our novel learning strategy and the associated registration method are shown to outperform other approaches. Theoretical, synthetic and real-world examples illustrate this improvement.

1. Introduction

Image registration, or geometrically aligning image volumes from different sources, is a fundamental problem for medical image analysis. The standard registration method seeks a transformation that aligns one floating image to a reference image such that the cost between the reference image and the transformed image is minimum. The complexity of the human brain and its natural variation between subjects often makes this optimization problem quite challenging, even when transformations are restricted to the affine space, as indicated by ongoing work in the field [20, 2, 12, 19, 17, 18].

Assuming a fixed cost function (here, the mutual information), the key aspect of this problem is the optimization strategy. A variety of methods based on local optimization have been used to optimize related similarity criteria [12, 19]. These methods include the gradient descent, Levenberg-Marquardt method, conjugate gradient descent, Newton's method and also non-gradient methods like Powell's method. Without any prior knowledge, the initial transform is often set as the identity transform. These methods are widely used in the domain of medical image analysis and give satisfactory results in many cases ([7, 17, 18]). However, due to the non-convex nature of the cost functions, these optimization methods are faced with the fundamental problem of local optima.

In order to estimate a global optimum and reduce the effect of the initial search position, methods like the multi-start method and simulated annealing, or, the coarse-to-fine multi-resolution search method [9, 4] have been proposed. They require a huge number of iterations to converge and are thus very time-consuming. These global optimization methods are designed for general optimization problems. However, ignoring knowledge about the specific problem requires randomly sampling huge numbers of transform parameters. This is an inefficient strategy when one considers that it is exactly prior knowledge that guarantees an algorithm's good performance on a specific data set (the No Free Lunch theorem [16]). This important fact was also noted by Ashburner et al. [1], who used a Gauss-Newton method to optimize affine registration in Bayesian framework. Their priors are directly modeled as Guassian distribution on the optimal transforms, which might have a large variance. Jenkinson, et al. also investigated restricting the transformation search space to sensible values and incorporating different tolerances and step sizes during iteration [9]. However, Jenkinson's approach used *ad-hoc* rules for determining these settings.

This paper focuses on how to use prior learning to achieve a better optimization strategy and thus gain improved registration results. Our strategy for improving upon standard affine registration methods is to automatically learn a *non-parametric* prior distribution of *potential transform improvements*. Given a gradient descent solution

978-1-4244-1631-8/07/$25.00 ©2007 IEEE

as initialization, we explicitly compare the gradient descent solution to the solution gained by the multi-start method. We define this gap between the global optimal transform and the transform of the local (gradient descent) minimum as the *tangent transform*. This gives a more constrained space than directly modeling the optimal transform as in [1]. Note that, in this study, this residual tangent transformation is also an affine transformation. A set of such tangent transforms from the training set enables one to learn the *tangent transform space*. This much smaller transformation space gives a tighter constraint on the affine transform that needs to be explored after a single gradient descent solution is gained.

We provide a theoretical argument and experimentally demonstrate that using a prior-based strategy within the restricted tangent transform space improves upon both gradient descent and general multi-start methods. We perform this analysis with respect to a common image registration task: registering every image in a database to a standard template. Typically, these images come from a single imaging protocol and a relatively homogeneous population. They therefore often share similar statistics, like possible scale and possible rotation angles. This restricted data set enables us to learn the "interesting" transformation space and largely constrain the optimization.

The remainder of this paper is organized as follows. Section 2 introduces the notion of the tangent transform and discusses how to learn its prior. Section 3 describes how to use the prior to aid the general multi-start method. Section 4 gives the experimental results of affine registration on a synthetic and a real database. Section 5 discusses the possible extension and the limitation of the work and concludes the whole paper.

2. Prior Learning on Tangent Transform

Our approach directly addresses the concern of how to learn a relevant parameter space within which to optimize a transformation. We assume that the cost function (eg, Mutual Information [15, 3]) is given and also a basic optimization procedure (eg, gradient descent) exists for that cost function. The basic local optimization method used here has been evaluated on brain images and applied extensively by the open-source ITK community [19, 8]. A subset of the fixed database will be used as the learning set to bootstrap the prior.

2.1. Cost Function and Its Optimization

Given a reference image $I^r(x)$ and another floating image $I^f(x)$, the aim of registration is to find the optimal transform T in the transformation space $\mathbf{S}_\mathrm{T}$. The definition of optimality is given by the cost function $C(I^r, I^f, T)$. The registration problem is formulated as finding the optimal T^* in space $\mathbf{S}_\mathrm{T}$ such that,

$$T^* = \arg\min_{T \in \mathbf{S}_\mathrm{T}} C(I^r(x), I^f(T(x)). \tag{1}$$

Although the challenging problem of defining a good cost function is out of the scope of this paper, the cost function is the vital key to the success of the registration. No optimization algorithm can obtain a good registration when the cost function is inappropriate. For this paper we use the widely accepted Mutual Information (MI) ([15, 3]) as our cost function. MI was first proposed to register multimodality images and is widely used in medical image registration ([10, 14, 12]). It measures the degree to which information from one image predicts another. This information-theoretic criterion does not assume specific intensity relations between two images and can be applied in different modalities. Consider image I and J as discrete random variables over intensities. The MI between I and J is defined as:

$$\begin{aligned} MI(I,J) &= \sum_i P_I(i)\log P_I(i) + \sum_j P_J(j)\log P_J(j) \\ &\quad - \sum_{i,j} P_{IJ}(i,j)\log P_{IJ}(i,j). \end{aligned} \tag{2}$$

$P_\cdot$ is the distribution for the random variable $(\cdot)$. P_{AB} is the joint distribution of A and B. Different implementations of MI typically vary the method of Parzen window estimation for the probabilities ([15] [14] [11]). The implementation of [11] is used for the experiments in this paper.

2.2. Transformation Space

The transformation space defines all the feasible transforms. The only requirement for the transform in our scheme is that it can be parameterized. For most databases, only a small portion of the full space contains optimal transformation parameters. Intuitively, one should not allow exploration of the full transform parameter space.

For this paper we restrict ourselves to the affine transform for the following reasons. First, it is the most flexible linear transform. Besides containing the rigid transform, it allows scale change in every dimension and shearing as well. Thus, it is suitable as a low-dimensional transformation for inter-subject studies. With 12 degrees of freedom, the prior parameters are not easy to set manually. Also a good affine registration is needed as a preprocessing step in non-rigid registration to make the global warping as small as possible. While we restrict this work to the affine space, the possibility of using our learning scheme with a larger deformable space is discussed in section 5.

The affine transform is represented as a projection matrix A and a translation vector t: $T(x) = \mathrm{A}x + t$. The projection matrix is decomposed into the product of the rotation matrix R, the scaling matrix S and the shearing matrix K, $\mathrm{A} = \mathrm{R} \times \mathrm{S} \times \mathrm{K}$.

Given the rotation axis of (u, v, w) and the rotation angle θ, the rotation is parameterized by the quaternion of four-parameters $\mathrm{R}_T = (a, b, c, d) \in \mathbb{R}^4$ with $(a, b, c, d) = (\cos\frac{\theta}{2}, u\sin\frac{\theta}{2}, v\sin\frac{\theta}{2}, w\sin\frac{\theta}{2})$. The unitary constraint $\|\mathrm{R}_T\|_2 = 1$ keeps the scale unchanged under rotation. The rotation matrix R is given by $\mathrm{R} =$

$$\begin{pmatrix} a^2+b^2-c^2-d^2 & 2bc-2ad & 2ac+2bd \\ 2ad+2bc & a^2-b^2+c^2-d^2 & 2cd-2ab \\ 2bd-2ac & 2ab+2cd & a^2-b^2-c^2+d^2 \end{pmatrix}.$$

Details about quaternions can be found in [13].

Let $\mathrm{S}_T = (s_1, s_2, s_3)$ represents the three scaling factor in 3 dimensions. The scaling matrix S is given by $\mathrm{S} = \begin{pmatrix} s_1 & 0 & 0 \\ 0 & s_2 & 0 \\ 0 & 0 & s_3 \end{pmatrix}$. Let $\mathrm{K}_T = (k_1, k_2, k_3)$ represent the shearing factors. The shearing matrix K is given by $\mathrm{K} = \begin{pmatrix} 1 & k_1 & k_2 \\ 0 & 1 & k_3 \\ 0 & 0 & 1 \end{pmatrix}$.

The transform T is thus parameterized by concatenating these parameters: $T = (\mathrm{R}_T, \mathrm{S}_T, \mathrm{K}_T, t) \in \mathbb{R}^{13}$. Such a parametrization gives each parameter a semantic meaning. The partial derivative of the transform to each parameter can be analytically obtained. Furthermore let $d\mathrm{R}(T_1, T_2) = \|R_{T_1} - R_{T_2}\|_2$. $d\mathrm{S}$, $d\mathrm{K}$, dt have similar definitions. We define the distance metric in parameter space as:

$$d(T_1, T_2) = \max(w_R d\mathrm{R}, w_S d\mathrm{S}, w_K d\mathrm{K}, w_t dt), \quad (3)$$

in which, w_R, w_S, w_K and w_t are preset weights.

The sequential composition of two transforms T_1 and T_2 is denoted as $T_2 \circ T_1$. The inverse transform of T is denoted as T^{-1}. QR decomposition is used to compute unique rotation, scaling and shearing parameters from A.

2.3. Tangent Transform

Most previous work seeks to enhance image registration's insensitivity to initialization in one or both of the following two ways. The first approach uses a multi-resolution strategy as in [14, 8]. The image at a coarse resolution has smoother features and fewer pixels and the cost function is easier to optimize. The spatial transform obtained at a coarser level is propagated to each successive level as the initialization. Another orthogonal approach puts a search boundary and step size in the transform space as in [9]. The values of the boundary and the step size are determined by human experts and must be tuned to each database. These two ways can be combined together. In this paper we target at improving in the second direction, which is to learn a more constrained space of possible transforms for the given database.

Directly learning the distribution of the optimal transform is not a good idea since it has nothing to do with the image itself. Instead, we propose the idea of the *tangent transform* to learn the possible range of the transform. The notion of the tangent transform is based on two observations. *First*, in a given brain image database, the possible T^* is almost always within a small subset of the full parameter space. So, it is only necessary to search the transform in a constrained space. *Second*, if the floating image does not have a dramatic variation from the reference image, the gradient descent T_g usually gives a reasonable solution estimate, but not a global optimal solution. Therefore if we can measure the gap between the global minimum T^* and the local minimum T_g, such a gap will be tighter than the gap between T^* and the identity transform T_{id} and thus the global minimum will be easier to find.

Based on this intuition, we formally define the notion of the tangent transform. Given the global optimal transform T^*, the *tangent transform* of T_g is defined as

$$\tilde{T}_g = T^* \circ T_g^{-\mathbf{1}}. \quad (4)$$

Since T_g is known, the optimization of T^* is equivalent as finding $\tilde{T}_g$. T^* is given by $T^* = \tilde{T}_g \circ T_g$

Another advantage of the tangent transform is that it is a *relative* measurement. T_g varies by different images. As long as T_g can capture the coarse pose variation, the small systematic pose change in T^* will minimally affect the prior distribution of $\tilde{T}_g$. Next we present our approach for estimating the distribution of the tangent transform.

2.4. Prior Learning

The ground truth of the global minimum T^* is impossible to get in many practical cases. However we still can get a sub optimal estimation from other methods, like the multi-start method discussed below. Normally these methods do not make any prior assumptions about the constraints in the transform parameter space. Instead, the parameter space is searched uniformly. But in fact many regions in the parameter space do not provide a reasonable registration. In addition, some areas are more important than others and need more careful exploration.

For a given training database, let T^*_m be the sub global minimum obtained by the general multi-start method. The practical tangent transform is computed as $\tilde{T}_g = T^*_m \circ T_g^{-1}$. Each parameter is viewed as a random variable. Figure 1 shows the histogram distribution of the 5th parameter s_1 for $\tilde{T}_g$ (a) and T^*_m (b). Figure 1(c) shows that the 10 parameters in $(\mathrm{R}_T, \mathrm{S}_T, \mathrm{K}_T)$ of T^*_m have smaller deviations than of $\tilde{T}_g$ except for the 10th. This shows the distribution of the tangent transform is distinct from uniform and more constrained than the space of T^*_m (as in [1]) and thus easier to sample.

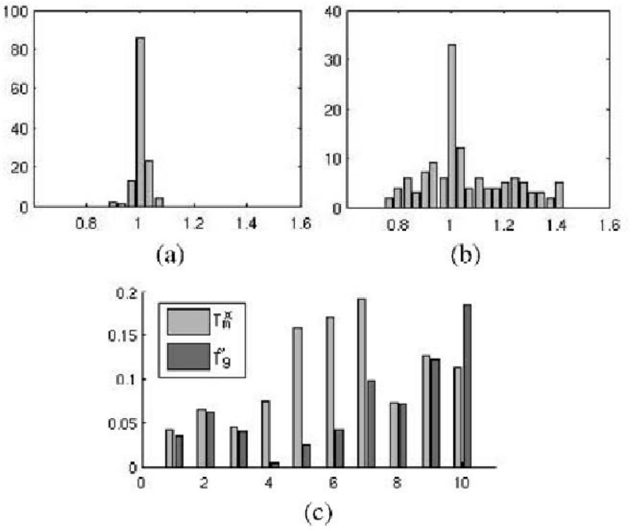

Figure 1. Illustration of the prior of transformation parameter. the histogram of s_1, the scaling factor on the x-axis, from (a) the tangent transform $\tilde{T}_g$, (b) the sub global optimum transform T_m^*. (c) comparison of standard deviation of the first 10 parameters (excluding the translation parameters). The first 4 are the rotation. The 5th to 7th are the scaling. The 8th to 10th are the shearing.

The learning technique helps to learn the values of the constrained space instead of setting them manually allowing our method to be easily extended to other datasets. This is an advantage as the learned space might not be easily guessed from experience (for example, the 10th parameter of $\tilde{T}_g$).

Theoretically we should learn the joint distribution of all the parameters. But since each parameter has its own semantic meaning and learning the joint distribution of 13 parameters (with the unitary constraint for the quaternion) is intractable, we only learn the marginal distribution for the affine transform. Section 5 discusses the potential ways to learn the parameters for non-rigid transform.

2.5. Simulation of Sampling Strategy

This section gives a theoretical argument for the efficacy of our sampling strategy. The multi-start method can be modeled as a sampling procedure. To register a database of N images to a template, the ground truth is regarded as a random variable from a distribution P_g. The multi-start samples the registration from another distribution P_q. For the general multi-start method, P_q is the uniform distribution. For our multi-start with prior, $P_q = P_g$. To simplify the analysis, P_g and P_q are assumed to be defined on the discrete space $\{1, \ldots, K\}$. For the ground truth $a \in \{1, \ldots, K\}$, the multi-start method samples a sequence of T possible answers $(s_1, \ldots, s_T)$ (with a little abuse of notation). $s_t = a$ for some $t \leq T$ means that the multi-start method finds the correct registration result at the t-th sample. In a realistic optimization this means that from the initial transform s_t the gradient descent (or other baseline iterative optimization) can converge to a when s_t and a are close enough.

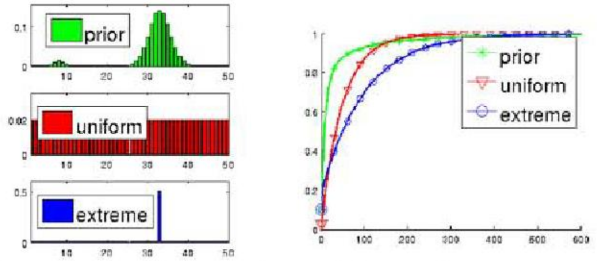

Figure 2. Monte-Carlo experiment of the sampling strategy. Left: three possible P_q. Right: $P(t \leq T)$ versus T. The ground truth is sampled $N = 50$ times from the *prior* distribution distribution for $K = 50$. For each ground truth, all the methods sampled at most 600 possible answers. $P(t \leq T)$ is averaged by repeating the experiment 100 times.

The performance of the registration can be modeled by a random variable t. $(t \leq T)$ means that the ground truth a is in the random sequence $(s_1, \ldots, s_T)$ of T elements. Note that a is a random variable of P_g. $P(t \leq T)$ is the probability of finding the correct answer by sampling T times. It is easy to derive that (see Appendix)

$$P(t \leq T) = 1 - \sum_{a=1}^{K} p_a (1 - q_a)^T, \qquad (5)$$

in which $p_a = P_g(a)$, $q_a = P_q(a)$.

To compare different P_q, we use Monte-Carlo experiment to simulate $P(t \leq T)$. Besides our prior scheme of $P_q = P_g$ and the general multi-start method of the uniform P_q, we also simulate a greedy strategy named *extreme*, which is a single peak distribution by sampling the mode of P_g for most time. Figure 2(b) shows the simulation result for a two-peak discretized distribution P_g. All curves converge to 1 for enough big T, which means that all the strategies find the correct answer by sampling enough points. However *prior* is significantly better than the other two strategies when T is small ($T < 100$ in this case). This validates our procedure for sampling the learned prior and shows that it has a larger chance of finding the global optimum given a relatively smaller number of samples. Note that it is possible to find the optimal P_q given P_g and T but we are not going to discuss it in this paper.

3. Algorithm

For this paper we use the multi-start algorithm as the baseline to get the sub global optimal transform to train the prior. Then we show how to incorporate the prior into the general multi-start method.

```
FOR $i = 1$ to MAX_ITER_NUMBER
    Random sample $T_0^i$ uniformly, such that
        $d(T_0^i, T^j) > d_{th}, \forall\, T^j \in$ SEARCH_HISTORY.
    Find local optimal solution $T^i$ by gradient descent from $T_0^i$,
        $T^i = \text{grad}(T_0^i, I^r, I^f)$.
    Compute the cost $c^i = C(I^r, I^f(T^i))$.
    Add $(T_0^i, T^i, c^i)$ into SEARCH_HISTORY.
END
Find the best solution $T^*$ from the history,
    $T_m^* = \arg_{T^i} \min c^i$, for $(T_0^i, T^i, c^i) \in$ SEARCH_HISTORY.
```

Table 1. Algorithm of general multi-start method.

3.1. General Multi-start Method

The multi-start method is one way to explore the transform parameter space. It is combined with other local optimization method. It restarts the search for the global optimum from a new solution once a region (or a path) has been explored by local optimization. The new start position is uniformly sampled from the transformation parameter space.

The pseudo-code for the multi-start procedure is illustrated in Table 1. A starting point T_0^i is constructed at iteration i, which is distinct from all the points searched in the history. The next step improves T_0^i by local gradient descent to a better solution T^i. The new solution T^i is added into the history record of the search list. When enough starting positions have been explored, the best result T^* in the history record is given as the final solution.

3.2. Learning the Prior Distribution of the Tangent Transform

On a given training set, the prior of the tangent transform is learned by computing the gap between T_m^* from the general multi-start method and T_g from the gradient descent. We assume that each parameter of the affine transform is independent. The prior is calculated as the marginal distribution from the training set of tangent transforms. To get a more robust counting, we use the *good* transforms that have a cost close to the best one found by the multi-start procedure.

Table 2 lists the pseudo code of learning the prior distribution P^i, the i-th parameter of the tangent transformation. Note that the same method could be used even if $T_g = T_{Id}$. However, more iterations of the multi-start method might be required to get an accurate estimate of the optimal transform, if T_{id} were chosen to compute $\tilde{T}_g$ instead of T_g.

3.3. Incorporating Prior with Multi-Start method

Once we have learned the marginal prior distribution for each parameter, we sample the tangent transform parameter according to the prior instead of according to uniform distribution. As mentioned previously, the prior gives a tighter constraint in the transform space and thus reduces the time required to find an improved solution.

```
Set $\mathbf{T} = \varnothing$
FOR the $j$-th image in the training set
    Recall SEARCH_HISTORY and $T^*$ for the $j$-th image.
    Compute the local minimum $T_g$ from the identity transform,
        $T_g = \text{grad}(T_{Id}, I^r, I^f)$.
    Collect the good tangent transform set $\mathbf{T_j}$,
    $\mathbf{T_j} = \{\tilde{T}_g | \tilde{T}_g = T \circ T_g^{-1}, C(T) < rC(T^*), T \in$ SEARCH_HISTORY$\}$,
        $C(T) = C(I^r, I^f, T), r = 0.8$.
    Set $\mathbf{T} \leftarrow \mathbf{T} \cup \mathbf{T_j}$
END
FOR the $i$-th transformation parameter
    Compute $P^i$ = the distribution of $\{a | a = T_0(i), T_0 \in \mathbf{T}\}$,
        $T_0(i)$ is the $i$-th element of $T_0$.
END
```

Table 2. Algorithm of learning the prior distribution of $\tilde{T}_g$.

```
Compute the local minimum $T_g$ from the identity transform,
    $T_g = \text{grad}(T_{Id}, I^r, I^f)$.
FOR $i = 1$ to MAX_ITER_NUMBER
    Random sample $T_0^i$ from the prior distribution $P$, such that
        $d(T_0^i, T^j) > d_{th}, \forall\, T^j \in$ SEARCH_HISTORY.
    Find local optimal solution $T^i$ by gradient descent from $T_0^i \circ T_g$,
        $T^i = \text{grad}(T_0^i \circ T_g, I^r, I^f)$.
    Compute the cost $c^i = C(I^r, I^f(T^i))$.
    Add $(T_0^i, T^i, c^i)$ into SEARCH_HISTORY.
END
    Find the best solution $T^*$ from the history,
    $T_m^* = (\arg_{T^i} \min c^i) \circ T_g$, for $(T_0^i, T^i, c^i) \in$ SEARCH_HISTORY.
```

Table 3. Algorithm of Incorporating prior with multi-start.

Similar to the training procedure, the multi-start method with prior begins by finding the local minimum T_g from the identity transform. To fill in the gap between T_g and the desired T^*, the potential tangent transform is sampled from the prior P and the gradient descent is performed from each sample. Finally, the transform with the best cost is chosen as the final solution. The pseudo-code to register I^f to I^r is listed in Table 3.

4. Results

To validate the performance of our algorithm, the first experiment was done on a synthetic database. One 3D T1 MRI image of a human brain was used as the template (in Figure 5). Random affine transforms were applied to the template image together with a small nonlinear random perturbation. 48 synthetic images of dimension 256×256×124 were generated altogether; examples are shown in Figure 3. We use the previously evaluated

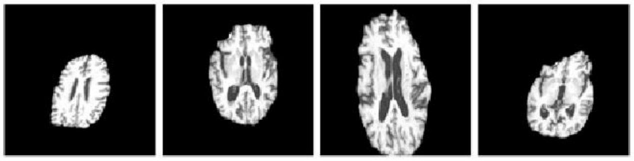

Figure 3. Examples of the synthetic 3D images.

Strategy $\mathcal{S}$	dR	dS	dK	dt	Iter #
grad 0	0.044	0.095	0.202	1.685	1000
mstart 200	0.025	0.094	0.156	1.612	500
prior 50	0.021	0.040	0.093	1.554	150
prior1 50	0.019	0.039	0.097	1.235	150

Table 4. Comparison of the transform parameters for different strategies on synthetic data. *prior 50* and *prior1 50* outperform *mstart 200*. The last column is the number of iterations for gradient descent in each strategy. For *prior 50* and *prior1 50*, 150 iterations are performed after the 1000 iterations of *grad 0*.

ITK implementation ([8, 11]) of mutual information as the cost function. Only the 10 parameters of $(\mathrm{R}_T, \mathrm{S}_T, \mathrm{K}_T)$ are learned as a non-parametric model.

The baseline method *grad 0* is T_g, the local gradient descent from the identity transform. This method was an extension to Lydia Ng's rigid transform registration, which was shown to perform well compared to similar methods in the Retrospective Image Registration Evaluation Project at Vanderbilt ([6]). The general multi-start method uniformly sampled 200 transforms as initial positions, denoted as *mstart 200*. *prior 50* and *prior1 50* are two tests of multi-start with prior, sampling 50 transforms in the tangent transform space. Each test randomly selected 24 images as the training set and the rest in the database as the test set. *grad 0* and *mstart 200* are tested on the whole database. The numbers of iterations of gradient descent for each strategy are listed in the last column of Table 4. Note that for *prior 50* and *prior1 50* the 150 iterations are performed after the 1000 iterations of *grad 0*.

The affine transforms for generating the synthetic data are used as the groundtruth of the transform for the registration. Table 4 gives the average distance of dR, dS, dK and dt for each method. By sampling from the prior learned from *mstart 200*, *prior* and *prior1* both outperformed other strategies. Note that the iterations of gradient descent in *prior* and *prior1* are less than in the general method *mstart 200*.

Three image metrics are computed to evaluate the performance of the registration: the MI, the mean square error (MSE) and the normalized correlation (NC) between the registered image and the template. Let I and J be the vector of the image intensity. MSE and NC are defined as:

$$MSE(I,J) = \|I-J\|_2^2, \quad NC(I,J) = \frac{\langle I,J \rangle}{\|I\|_2 \|J\|_2}.$$

Strategy $\mathcal{S}$	rMI	$rMSE$	rNC	mNC
grad 0	0.00	0.00	0.00	94.38
mstart 200	2.01	6.58	0.66	94.97
prior 50	4.57	14.9	1.38	95.29
prior1 50	3.62	13.4	0.88	95.48

Table 5. Evaluation of different strategies on synthetic data. The data in 2nd to 4th column is shown in percentage (%). *prior 50* and *prior1 50* outperform *mstart 200*.

With *grad 0* as a baseline, the relative increase of MI (rMI) and NC (rNC) and the relative decrease of MSE (rMSE) are averaged over the database. For the optimization strategy $\mathcal{S}$ and the corresponding registration result $T^{\mathcal{S}}$, they are defined as:

$$\begin{aligned} rMI(\mathcal{S}) &= \frac{1}{N}\sum_{i=1}^{N} \frac{MI(I^r, I_i^f(T_i^{\mathcal{S}}))}{MI(I^r, I_i^f(T_{i,g}))} - 1, \\ rMSE(\mathcal{S}) &= 1 - \frac{1}{N}\sum_{i=1}^{N} \frac{MSE(I^r, I_i^f(T_i^{\mathcal{S}}))}{MSE(I^r, I_i^f(T_{i,g}))}, \\ rNC(\mathcal{S}) &= \frac{1}{N}\sum_{i=1}^{N} \frac{NC(I^r, I_i^f(T_i^{\mathcal{S}}))}{NC(I^r, I_i^f(T_{i,g}))} - 1. \end{aligned}$$

The average of the normalized correlation (mNC) is also reported:

$$mNC(\mathcal{S}) = \frac{1}{N}\sum_{i=1}^{N} NC(I^r, I_i^f(T_i^{\mathcal{S}})).$$

These image metrics on the synthetic data are summarized in Table 4. The sub-global optimal method of *mstart 200* gives 2% increase of rMI compared with the gradient descent. After learning the tangent transform prior from *mstart 200*, the performance of *prior 50* and *prior1 50* has around 4% improvement by only 50 samples, which is 25% of *mstart 200*.

To evaluate performance on real-world data, we tested the algorithm on a database of 67 T1structural MRI images. The images are from elderly and neurodegenerative human brains, collected from a 1.5T Siemens scanner at 1×1×1.5mm resolution. The dimension of each 3D image is 256×256×124. All images are preprocessed to remove skulls before registration.

Each of *prior 50* and *prior1 50* used a randomly selected set of 33 images as the training set and the rest as the test set. *grad 0* and *mstart 200* are tested on the whole database. The numbers of iterations for each initial transform are listed in the last column of Table 6. As there is no groud truth for the affine transform, only the evaluation on the image metrics was reported. The average metric valuess are summarized in Table 6. Figure 4 shows the histogram of rMI and $rMSE$ for *mstart 200* and *prior 50*.

mstart 200 only gives 1% increase of rMI compared with the gradient descent. This shows that the gradient descent works for many image registration cases. However

Strategy S	rMI	$rMSE$	rNC	mNC	Iter #
rand 2K	-20.87	-84.01	-13.05	74.05	N/A
rand 2K+trans	-17.69	-51.54	-9.13	77.39	N/A
grad 0	0.00	0.00	0.00	92.29	1000
mstart 200	0.98	7.04	0.61	92.97	500
prior 50	3.11	14.10	1.24	93.67	150
prior1 50	2.99	11.60	1.13	93.44	150

Table 6. Evaluation for different strategies on the real data. *prior 50* and *prior1 50* outperform *mstart 200*. The last column is the number of iterations for gradient descent. For *prior 50 / prior1 50*, 150 iterations are performed after 1000 iterations of *grad 0*. The 2nd to 4th columns are shown in percentage (%).

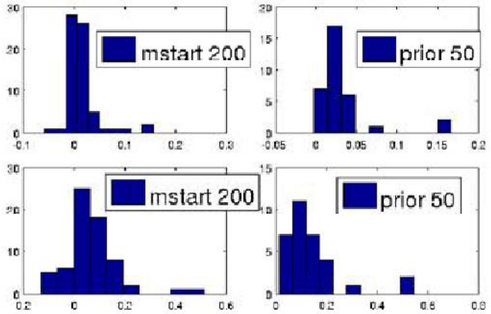

Figure 4. Histogram of *rMI* (1st row) and *rMSE* (2nd row) for *mstart 200* (on left) and *prior 50* (on right).

there is still space to improve. After learning the tangent transform prior from *mstart 200*, the performance of *prior 50* and *prior1 50* give a 3% improvement by 50 samples. The prior learning again outperforms the sub-global optimal from *mstart 200* with less computational overhead.

An interesting comparison is to randomly sample 2000 parameters without any gradient descent optimization (*rand 2K*). This gives a 20.87% decrease of rMI in the performance. If the gradient descent is only allowed on the translation parameters (*rand 2K+trans*), there is still 17.69% decrease. This shows that gradient descent on all the parameters is necessary. It is not sufficient to do random sampling without further optimization.

Finally, we emphasize that the new prior-based affine registration method not only produces better metric values, but also produces visible improvements in the results. As shown in Figure 5, both *prior 50* and *mstart 200* give a better result than *grad 0*. At the same time, the overall brain shape is registered better by *prior 50* than by *mstart 200*. This example has $rMSE = 16.7\%$, which is representative of the average improvement gained by our methods as indicated by the histogram in Figure 4.

5. Conclusion and Discussion

In this paper, we proposed the notion of the tangent transform which represents the gap between a global optimum and a locally optimal (or fixed) solution. We used the tangent transform to learn prior distributions of transform parameters in a restricted space. We performed a careful analysis and showed that the learned prior of the tangent transform helped to explore the relevant affine transform space. Our learning method has the following benefits:

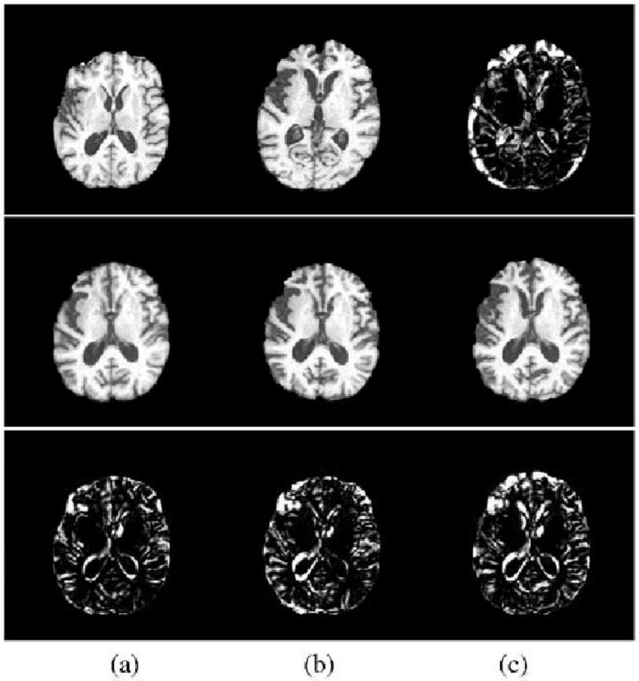

Figure 5. Comparison of registration results, shown in the transverse view. 1st row: (a) the template, (b) the moving image, (c) the square differencing image of (a) and (b). 2nd row: the registration results from (a) *prior 50*, (b) *mstart 200*, (c) *grad 0*. 3rd row: the square differencing image for 2nd row with the template. The *rMSE* of *prior 50* is 16.7% for this image.

- There is no restriction or Gaussian assumption on the parameter space. The algorithm will learn the non-parametric distribution to save the labor of human experts setting new searching rules. It therefore automatically extends to new datasets.
- There is no need to provide the ground truth for training. The algorithm can learn the prior from the general sub-global optimization method. The results show that our algorithm surpasses the performance of the general sub-global optimization with the learned prior.
- Both theoretical simulation and real data experiments show that it uses fewer iterations to get better performance than the general multi-start method.

Our notion of the tangent transform defined a practical prior model by explicitly modeling the residual gap between the estimated global optimal solution and the gradient descent solution. This is different from modeling the optimal solution directly as in [1]. Even in the case when the baseline gradient descent did a reasonable registration, this technique still leads to improvements.

The core assumption of the method is that all the images in the same database have a rough and learn-able range of variation. This is validated for our database by the evaluation. However this method may fail if some rare images have an outlier tangent transform distortion compared with other images. The learned prior may not work on this case.

The framework proposed in section 3.3 is both simple and extensible. First it is independent from the choice of cost function. Second this framework is easily incorporated with other general random optimization procedures, like simulated annealing.

The transform used in this paper is the affine transform. It is also possible to extend the work to non-rigid transforms, like the B-spline or thin-plate spline transforms. In these cases there will be many more parameters to learn. There are two possible ways to handle this issue. The first is to learn the joint distribution by dimension reduction techniques, like Principal Component Analysis. The second is to generate synthetic data by randomly warping images.

This framework can also be extended in a way of Sequential Monte Carlo methods ([5]). In the terms of Bayesian model, the registration transform is the hidden variable and the learned prior of the tangent transform is the observation model. The prior of tangent transform is defined as the gap between the global optimal transform and the local optimal transform starting from the identity transform. A second level of tangent transform can be defined as the gap between the global optimal transform and the current optimization. As each level leads to a more and more constrained transform space, the optimization can be done with fewer and fewer samples and lead to well-defined convergence.

The overall performance gains indicated by this approach were consistent across theory, simulated data and real data. We believe this indicates the strength of the concepts and shows that even simple non-parametric learning strategies may be very useful in improving registration methods.

Acknowledgement The authors gratefully acknowledge NIH support of this work via grants EB006266, NS045839, MH073529 and MH068066.

References

[1] J. Ashburner, P. Neelin, D. Collins, A. Evans, and K. Friston. Incorporating prior knowledge into image registration. *NeuroImage*, 6(4):344–352, 1997. 1, 2, 3, 7

[2] L. G. Brown. A survey of image registration techniques. *ACM Computing Survey*, 24(4):325–376, 1992. 1

[3] A. Collignon, F. Maes, D. Delaere, D. Vandermeulen, P. Suetens, and G. Marchal. Automated multimodality image registration using information theory. *Proc. of Information Processing in Medical Imaging*, pages 263–274, June 1995. 2

[4] O. Cordon and S. Damas. Image registration with iterated local search. *J. Heuristics*, pages 73–94, 12(2006). 1

[5] A. Doucet, J. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods In Practice*. Springer-Verlag, 2001. 8

[6] J. M. Fitzpatrick. *The Retrospective Image Registration Evaluation Project*. http://www.vuse.vanderbilt.edu /~image/registration/reg_eval_html/Ng.html. 6

[7] L. Ibáñez, L. Ng, J. C. Gee, and S. R. Aylward. Registration patterns: the generic framework for image registration of the insight toolkit. In *ISBI'02*, pages 345–348, 2002. 1

[8] L. Ibáñez, W. Schroeder, L. Ng, and J. Cates. *The ITK Software Guide*. Kitware, Inc., 2nd edition, 2005. 2, 3, 6

[9] M. Jenkinson and S. Smith. A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, pages 143–156, May 2001. 1, 3

[10] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE Trans. Med. Imaging*, 16(2):187–198, 1997. 2

[11] D. Mattes, D. Haynor, H. Vesselle, T. Lewellen, and W. Eubank. Non-rigid multi-modality image registration. *Medical Imaging 2001: Image Processing*, pages 1609–1620, 2001. 2, 6

[12] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever. Mutual information based registration of medical images: A survey. *IEEE Trans. Med. Imaging*, 22(8):986–1004, 2003. 1, 2

[13] K. Shoemake. Animating rotation with quaternion curves. In *SIGGRAPH'85*, pages 245–254, New York, NY, USA, 1985. ACM Press. 3

[14] P. Thévenaz and M. Unser. Optimization of mutual information for multiresolution image registration. *IEEE Trans. Image Processing*, 9(12):2083–2099, 2000. 2, 3

[15] P. A. Viola and W. M. Wells III. Alignment by maximization of mutual information. In *ICCV'95*, pages 16–23, Washington, DC, USA, 1995. 2

[16] D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE Trans. Evolutionary Computation*, pages 67–82, April 1(1997). 1

[17] R. P. Woods, S. T. Grafton, C. J. Holmes, S. R. Cherry, and J. C. Mazziotta. Automated image registration: I. general methods and intrasubject, intramodality validation. *J. Comput. Assist. Tomogr.*, 22(1):139–152, 1998. 1

[18] R. P. Woods, S. T. Grafton, J. D. G. Watson, N. L. Sicotte, and J. C. Mazziotta. Automated image registration. ii. intersubject validation of linear and nonlinear models. *J. Comput. Assist. Tomogr.*, 22(1):153–165, 1998. 1

[19] T. S. Yoo, editor. *Insight into Images: Principles and Practices for Segmentation, Registration, and Image Analysis*. A K Peters Ltd., 2004. 1, 2

[20] J. Zhang and A. Rangarajan. Affine image registration using a new information metric. In *CVPR'04*, pages 848–855, Los Alamitos, CA, USA, 2004. 1

Appendix Derivation of Eqn 5:

$$P(t \leq T) = 1 - P(t > T) = 1 - \textstyle\sum_{a=1}^{K} P(a, t > T)$$
$$= 1 - \textstyle\sum_{a=1}^{K} P_g(a) P(s_1 \neq a, \ldots, s_T \neq a)$$
$$= 1 - \textstyle\sum_{a=1}^{K} p_a \prod_{t=1}^{T} P(s_t \neq a)$$
$$= 1 - \textstyle\sum_{a=1}^{K} p_a (1 - q_a)^T$$