# Efficient population registration of 3D data

Lilla Zöllei[1], Erik Learned-Miller[2], Eric Grimson[1], William Wells[1,3]

[1]Computer Science and Artificial Intelligence Lab, MIT;
[2]Dept. of Computer Science, University of Massachusetts, Amherst;
[3]Radiology Department, Brigham and Women's Hospital

## Abstract

We present a population registration framework that acts on large collections or *populations* of data volumes. The data alignment procedure runs in a simultaneous fashion, with every member of the population approaching the *central tendency* of the collection at the same time. Such a mechanism eliminates the need for selecting a particular reference frame *a priori*, resulting in a non-biased estimate of a digital atlas. Our algorithm adopts an affine *congealing* framework with an information theoretic objective function and is optimized via a gradient-based stochastic approximation process embedded in a multi-resolution setting. We present experimental results on both synthetic and real images.

## 1. Introduction and Motivation

The registration of two data sets is the problem of identifying a geometric transformation which maps the coordinate system of one to that of another or, more generally, establishing a homology among the input images when the number of input images to be aligned is more than two. In this scenario, it is not one, but a group of transformations that needs to be identified in order to put all the inputs into correspondence. We are particularly motivated by the *population* registration problem, which includes the registration of collections of images or volumes, where the number of inputs is greater than twenty or potentially much greater.

Depending on the nature of the images to be processed, we distinguish between mono- and multi-modal registration tasks. In the former, the inputs are acquired by the same, and in the latter case by different, types of imaging devices. A registration problem that lies between these two categories is the so-called template-to-subject registration which involves the alignment of an image or a set of images with a template constructed independently from the current alignment. (The template, as explained below, can be one member of the input image set or a probabilistic representation of prior knowledge about the imaged object.) In computer vision, template-registration tasks are common when there is some prior information about the standard characteristics of the input and / or when one wishes to compare a current sample of a group to previously processed ones. The results can then be further studied to carry out statistical inference on shape, population characteristics or on abnormal variability. They could also be used as a pre-processing step for segmentation studies.

In the medical domain the same task has become increasingly important and is referred to as atlas-to-subject registration. Its prevalence can be explained by the accessibility of rapidly growing image databases and faster computers that allow for population studies and various data mining tasks. We were initially inspired by the availability of such data volumes; thus our examples are all from the medical domain. Note, however, that the algorithm formulation is very general and it is not restricted to only medical input data sets.

In this work, we demonstrate a new unbiased and computationally efficient framework for aligning populations of 3D medical images for the purpose of digital (anatomical) atlas construction. We believe that defining a robust inter-subject registration technology that enables the comparison of large numbers of images will allow us to build better structural atlases, and to further analyze inter-subject differences.

## 2. Background and Previous Work

Several approaches exist that propose the alignment of multiple data sets into the same coordinate frame. Besides the details of the registration algorithm applied, there is a significant difference in how each method interprets the common coordinate frame (or template). For some specific applications, the desired template is already established. The input volumes then do not need to be managed as a set, they can be aligned with the reference frame individually. This approach is advantageous when single input volumes need to be compared to the template. If, however, the input volumes are to be treated (simultaneously) as a group, other mechanisms are required.

For the rest of the applications, the digital template is not available, so that too has to be generated along with the aligning transformations. In the medical community, recently, there have been several approaches proposed [3, 6, 12, 11, 14]. One group of algorithms selects a standard coordinate frame (for example, based upon certain anatomical structures) and requires the algorithm to position all the inputs into that frame. The mean of the so-aligned images is then computed. Such methods have been performed, for instance, with the usage of the Talairach anatomical coordinate system [1, 13]. A major disadvantage of these methods is that the images need to be pre-processed in order to have the matching landmarks reliably identified in them. That is a time-consuming and potentially error-prone procedure.

Other approaches select one of the current data volumes to be the common reference frame [3]. After all the other volumes are aligned to this, their *mean* is computed. The problem here is the introduction of bias into the procedure by claiming that one sample volume can represent the standard reference. Even if the procedure is re-run several times or the selection of the particular reference frame is carried out in a more careful manner, we cannot always ensure a non-biased implementation of this process. In the case of anomalies present in the input, the registration results could be significantly distorted.

Instead, there is growing interest in generating *mean models* as a *by-product* of a larger-scale registration process. That formulation eliminates the introduc-

tion of a bias into the registration framework by simultaneously evolving the data sets towards a common reference. According to one approach, the "mean" is initially defined and the images are aligned to that reference image [5]. The process is iterated until the optimal alignment is found. Another approach follows that same scheme, but it performs non-rigid alignment of 2D scans using a minimum description length criterion [12]. Because of memory limitations, these algorithms can currently handle only a limited number ($< 10$) of input volumes. We note that algorithms in this subgroup are closely related to a maximum likelihood framework where each voxel distribution is represented by a Gaussian with a mean equal to the voxel mean and a fixed variance. When using our framework - congealing - though, each voxel has a separate, individually optimized non-parametric distribution. Since the distribution of tissues at a particular voxel is usually highly non-Gaussian, it would seem that our framework is more appropriate.

Another approach within this same category defines the image set registration problem by the generalization of a one-to-one alignment framework [10]. The authors estimate the joint density function of all the inputs and construct a maximum likelihood-type similarity metric. For computational ease the input images are pre-segmented into a handful of anatomical classes. A drawback of this approach is that that it requires the construction of a joint density function whose size grows exponentially with the number of input images. While the amount of data available only grows linearly, the number of samples required for a good density estimation grows exponentially.

## 3. Our Method

We are interested in formulating the problem as an analogy to an inter-subject image set alignment task. We use a technique called *congealing* as a basis of our framework. This approach was first introduced in the machine learning and computer vision literature, offering a solution to the hand-written digit recognition problem [8, 9]. There, a model of the *central tendency* of binary input images was recovered and used for classification purposes.

The objective function proposed in the congealing framework is the total voxel-wise entropy of the input image volumes. The entropies are computed at each coordinate location and then these quantities are added together. This formulation thus models distributions of each voxel *conditioned on spatial location* rather than treating each position as equivalent. This is in contrast with the popular mutual information or joint entropy methods for alignment where entropy is measured *within* an image and the voxel distribution is assumed to be *i.i.d.*([7, 10, 15]). The sum of voxel-wise entropies is approximately equivalent to finding the maximum likelihood latent image in the population [8], and using it as an alignment criterion results in a low total entropy joint image. This outcome represents the underlying shape of the imaged objects and its residual variation.

Warfield et al. have already applied a preliminary version of the congealing approach to the problem of fusing MRI scans of 22 pre-term infants and producing an atlas of the developing white matter [14]. In that implementation, the

intra-cranial cavity (ICC) of all the input volumes was pre-segmented to allow for binary congealing, and one member of the population was also set to be stationary (resulting in a biased result). A nine parameter affine transformation was identified for all the inputs. (A model created by this method on adult brain scans is referred to as *control model* in Section 5 and is shown in Fig. 4 (b).)

Our contribution to the congealing framework lies in its adaptation to a population of grayscale-valued 3D data volumes without introducing any bias and a computationally efficient implementation via a stochastic gradient-based optimization procedure in a multi-resolution framework.

### 3.1. The Objective Function

As mentioned already, our congealing framework adopts the sum of voxel-wise entropies as a joint alignment criterion. The main intuition behind using such an objective function is that, when in proper alignment, intensity values at corresponding coordinate locations from all the inputs form a low entropy distribution. That statement holds even if the intensity values are not identical. Hence noise or bias fields, and what is more, corresponding multi-modal inputs can also be accommodated. An entropy-based objective function is also appropriate to handle data sets whose intensities form multi-modal distributions. That property is of great benefit when the population consists of (sufficient number of representatives of) data volumes with widely varying intensity profiles. For example, the tissue intensities at a particular voxel location in the cortex would likely include some white matter voxels, some gray matter voxels, and a small percentage of other tissue types. The distribution of brightness values in such a distribution is frequently multi-modal.

If we denote the collection of $m$ input volumes as $\mathcal{I} := \{I_1, I_2, ..., I_m\}$, then our goal is to identify the set of $m$ transformations, $\mathcal{T} := \{T_1, T_2, ..., T_m\}$ (one transformation associated with each volume), such that the objective function $f$ of total voxel-wise entropies is minimized. The objective function is then:

$$f(\mathcal{I}, \mathcal{T}) = f(T_1(I_1), ..., T_m(I_m)) = \sum_{i=1}^{N} H(\mathcal{I}(\mathcal{T}(\mathbf{x_i}))),$$

where $\mathbf{x_i} \in \mathcal{R}^3$ indicates a particular coordinate location in the data coordinate system, $H$ is the Shannon entropy and $N$ is the total number of voxel locations in the data coordinate system. This measure actually forms an upper bound on the true entropy of the image distribution. By minimizing this upper bound, we approximate the minimum of the true entropy [8].

In the current implementation we use 12-parameter affine transformations. Our convention orders the transformation components as the rotation, scaling and shearing followed by the displacement. Accordingly, $\forall j \; T_j(\mathbf{x_i}) = (D_j + Sh_j S_j(R_j(\mathbf{x_i})))$, where $D_j$ is the displacement, $R_j$ is the rotation, $S_j$ is the anisotropic scaling and $Sh_j$ is the shearing component of transformation $T_j$.

As both the size and number of our expected image volumes are large, memory allocation and computational speed are both of serious concern. Con-

sequently, we apply a stochastic sampling framework and the EMMA[1]-style entropy estimator in our framework [15]. Instead of considering all the locations in the data coordinate space, we propose a random selection of them. Then an approximation of the total sum of voxel-wise entropies is computed for a particular alignment configuration. We write the modified objective function (approximating expectation with sample average) as:

$$f(\mathcal{I}, \mathcal{T}) = -\frac{1}{m} \sum_{i=1}^{M} \sum_{j=1}^{m} \log p(I_j(T_j(\mathbf{x_i}))),$$

where $M$ now indicates the number of randomly selected sample points. Note, that the samples in this reduced set of coordinate locations are not fixed but re-generated at each iteration of the algorithm. As the experiments show, this modification enabled us to significantly reduce the overall number of voxel locations considered in our computations.

### 3.2. The Optimization

In the original framework of the congealing algorithm, a coordinate descent optimization was used to guide the minimization of the objective function. As this technique is not computationally efficient for our purposes, we have implemented an iterated stochastic gradient-based update mechanism (similar to that of [15]) that significantly reduces the processing time.

### 3.3. Transformation Normalization

We have a normalization step included at the end of each iteration, where we compose each transformation estimate by the inverse of the mean transformation matrices. This update is necessary as it ensures that the average movement of points at corresponding coordinate locations is zero, thus preventing the images from drifting out of the field of view.[2]

### 3.4. The Multi-Resolution Framework

It is widely known in the registration literature that optimization functions can easily become trapped in local minima. Although congealing already mitigates some problems of local minima [8], we also constructed a multi-resolution registration framework. This implementation starts the processing of the data sets at a down-sampled and smoothed level and then refines the results during the higher resolution iterations. Not only does this framework improve the optimization, it also boosts computation speed and memory usage efficiency. The number of hierarchy levels is mostly dependent on the quality and the original size of the input images. For the experiments presented in this work, it was sufficient to use a maximum of three levels of hierarchy.

---

[1] The name EMMA refers to "Empirical entropy manipulation and analysis"

[2] This normalization criterion is different from the one presented in [8], where the normalization aimed to maintain a zero mean displacement estimate and a mean transformation matrix of determinant 1.

## 4. Medical MRI Experiments

We ran experiments on three different populations of MRI acquisitions. The first set consisted of 22 baby brain volumes. Each brain volume was 176 by 186 by 110 voxels, with each voxel measuring 1.0 by 1.0 by 2.0 millimeters in size. The second and third data sets consisted of 28 and 127 adult brain volumes. These volumes were 256 by 256 by 124 voxels, with each voxel measuring 0.9375 by 0.9375 by 1.5 millimeters. Due to page limitations, we will demonstrate the results only on the third set of the images. We believe that this is the first report of simultaneous registration run on such a large collection of input volumes.

The experiments on the 127 medical scans were executed on three different resolution levels (where the volumes were (32 by 32 by 31), (64 by 64 by 62) and (128 by 128 by 124) voxels). The largest offset was obtained on the lowest level and then refinement was computed on the higher hierarchy levels. In our experiments we only had to select between 800 - 1500 samples, which constitutes just .05-2.5% of the total voxels, and no more than 250 iterations were necessary. The total running time for the experiment was approximately six hours.

The results of the experiments are displayed in Fig. 1 (a). This figure portrays three orthogonal slices of the mean volumes computed before and after the experiments. As a qualitative measure, we can establish that following the population alignment, the data volumes properly line up and the mean volumes have clean and sharp boundaries.

## 5. Validation

Validating our results and verifying our alignment is a complex task. In this section we provide both qualitative and preliminary quantitative results.

Visually we can confirm that the mean volumes computed after the congealing process have much sharper boundaries than prior to alignment (see Fig. 1 and Fig. 4 (c)). This is an indirect indicator of how good an agreement has been achieved. Looking at the central slices extracted from all the input volumes after the congealing process (see Fig. 2 (b) and 3 (b)) also suggests that the algorithm has managed to find a good quality alignment.

We also provide a quantitative analysis obtained from running our algorithm both on a synthetic image population and from comparing one of our adult brain models to an already existing one.

### 5.1. Synthetic Example

As a control study, we selected one particular medical MRI volume from a group of adult brain acquisitions and created a database of transformed volumes by applying affine transformations to it. The magnitude of these transformations varied between $+/-10$ degrees for rotation, $+/-10$ mm for displacement, between $[.85, 1.15]$ factors for scaling and between $+/-.1$ factors of shearing. At the onset of the algorithm, 40 volumes were randomly generated as inputs. All the input volumes were 124 by 256 by 256 voxels, with each voxel measuring

.9375 by .9375 by 1.5 mm. The twelve parameters of the affine transformations were recovered after running our algorithm on two levels of the hierarchy. The number of samples used was .05% of the total number of voxels and fewer than 400 iterations were necessary to achieve convergence. The total running time was 2964 seconds. The results of these experiments can be seen in Fig. 1 (b) and 2 (b). The former illustrates the mean volumes computed before and after the congealing process, while the latter displays the central slices of each of the input volumes before and after the alignment. For the initially selected adult brain scan, we had access to the segmentation of two sub-cortical structures, the left and right thalamus (LT and RT). After the congealing alignment was executed, we applied the resulting transformations to these segmentations and then computed an overlap measure on the so-aligned binary images. The measure of our choice was $f_{\text{overlap}}(A_1, A_2) = \frac{|A_1 \bigcap A_2|}{min(|A_1|,|A_2|)}$ ($A_i$ indicating binary variables), which can be easily generalized to higher number of inputs.

The overlap scores indicate great improvement, they increased from 0 to .745 and to .75 in the case of LT and RT, respectively. These numbers might seem a bit low, but as the overlap metric we use is quite conservative, we further interpret these results. For the left thalamus, .745 means that all 40 input segmentations agreed 74.5% of the time and 34 inputs are sufficient to reach an 89% score. Similarly, for the right thalamus, .75 means that all 40 inputs agreed 75% of the time, and 35 inputs are sufficient to reach a 90% score.

Several factors may influence the magnitude in a decrease of this score. First, when computing the intersection, even single misaligned voxels can significantly reduce the metric value. Second, transforming the binary structures introduces quite a high variation in the size of these relatively small anatomical structures: the standard deviation of the structure sizes (after the transformations have been applied with nearest neighbor interpolation) was 149 voxels.[3]

We also analyzed the transformations resulting from the congealing process. Computing exact error measurements (even when knowing the ground truth offsetting transformations) is difficult as the transformations recovered by our alignment process are able to recover the inverse of the offsetting transformations only up to a common term. Therefore, we recovered both a dispersion and a bias term of the resulting errors across all the input volumes via an analysis similar to the consistency measures introduced in [4]. Our dispersion scores (indicating accuracy) were in the range $[0.05, 0.15]$ and the bias terms (indicating the magnitude of the common term) in the $[0, 2]$ voxel range.

### 5.2. Atlas Comparison

As an additional experiment, we also compared one of our resulting atlases to a previously generated template. More specifically, we ran our algorithm on 22 adult brain volumes (with the same parameters as indicated in Section 5.1) and

[3] We indeed experimented with other interpolation methods, which resulted in lower standard deviations, but as the minimum component size was also increased in this manner, the end result did not change significantly from the one that we report here.

compared that to a *control model* whose generation is explained in details [14]. Qualitatively, we first assess the success of the congealing algorithm (Fig.3) and then we compare the atlases in Fig.4 (b) and (c) and establish that they are highly similar. (Note the 3D view of the mean volume in the original setup is demonstrated in Fig. 4 (a)).

For a quantitative analysis, we used the same segmentation-overlap study as in the case of the synthetic experiments. In the case of LT, our overlap measure was .474027 vs .428483 of the *control model* and in the case of RT we obtained .439664 vs .496284. Our performance thus is comparable to that of the atlas.

These overlap measures are even lower than in Section 5.1. That is because in this experiment we process inter-subject scans and the normal variability in their differences can only be explained to a certain extent by affine transformations. Currently we are implementing a viscous fluid-based non-rigid warp [2] to add to our multi-resolution framework. Such a dense deformation model should be able to eliminate some of the remaining local disagreements in our alignment results.

## 6. Summary and Conclusions

In this paper, we introduced a new population registration framework. Without any pre-processing step, we used a congealing-type alignment method to efficiently put a large collection of data volumes into correspondence. The algorithm builds on an information theoretic objective function and currently uses fully parameterized affine transformations. We introduced an approximate stochastic sampling framework which allowed us to process only a small number of samples from the inputs. The optimization is implemented in a stochastic gradient-based optimization framework that enables a substantial increase in speed.

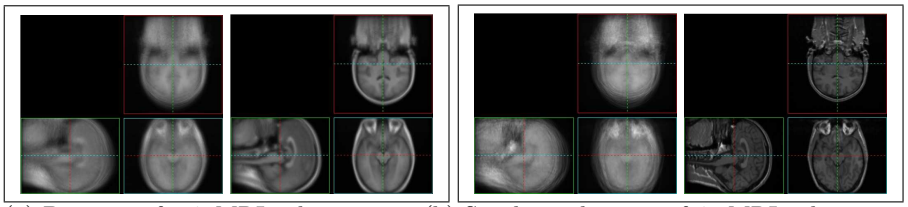
## 7. Acknowledgment

This work has been supported by NIH 5 P41 RR13218 and NIH Roadmap for Medical Research, Grant U54 EB005149 as a member of the National Alliance for Medical Image Computing (NAMIC). Information on the National Centers for Biomedical Computing can be obtained from http://nihroadmap.nih.gov/bioinformatics.

## References

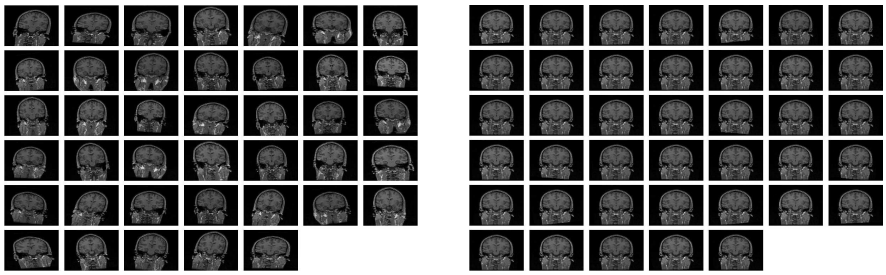
1. D. Collins. *3D Model-Based Segmentation of Individual Brain Structures from Magnetic Resonance Imaging Data*. PhD thesis, McGill University, Montreal, Canada, 1994.
2. E. D'Agostino, F. Maes, D. Vandermeulen, and S. P. A viscous fluid model for multimodal non-rigid image registration using mutual information. *Medical Image Analysis*, 7:565–575, 2003.
3. A. Guimond, J. Meunier, and T. J.-P. Average brain models: A convergence study. Technical Report 3731, INRIA, July 1999.

4. H. Johnson and G. E. Christensen. Consistent landmark and intensity-based image registration. *IEEE Transactions on Medical Imaging*, 21(5):450–461, May 2002.
5. S. Joshi, B. Davis, M. Jomier, and G. Gerig. Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage*, September 2004.
6. P. Lorenzen, B. Davis, and G. Gerig. Multi-class posterior atlas formation via unbiased kullback-leibler template estimation. In *LNCS*, volume 3216, pages 95–102, September 2004.
7. F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging*, 16(2):187–198, 1997.
8. E. Miller. *Learning from One Example in Machine Vision by Sharing Probability Densities*. PhD thesis, Massachusetts Institute of Technology, February 2002.
9. E. Miller, N. Matsakis, and P. Viola. Learning from one example through shared densities on transforms. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 464–471, 2000.
10. C. Studholme and V. Cardenas. A template-free approach to volumetric spatial normalization of brain anatomy. *Pattern Recognition Letters*, 25(10):1191–1202, July 2004.
11. P. Thompson, R. Woods, M. Mega, and A. Toga. Mathematical/computational challenges in creating deformable and probabilistic atlases of the human brain. In *Human Brain Mapping*, volume 9 (2), pages 81–92, February 2000.
12. C. Twining and C. Marsland, S. Taylor. Groupwise non-rigid registration: The minimum description length approach. In *Proceedings of BMVC*, 2004.
13. D. Van Essen, H. Drury, S. Joshi, and M. Miller. Functional and structural mapping of human cerebral cortex: Solutions are in the surfaces. In *National Academy of Sciences*, volume 95, pages 788–795, 1998.
14. S. Warfield, J. Rexilius, P. Huppi, T. Inder, E. Miller, W. Wells, G. Zientara, F. Jolesz, and R. Kikinis. A binary entropy measure to assess nonrigid registration algorithms. In *Fourth International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Lecture Notes in Computer Science, pages 266–274. Springer, October 2001.
15. W. Wells, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis. Multi-modal volume registration by maximization of mutual information. In *Medical Image Analysis*, volume 1, pages 35–52, 1996.



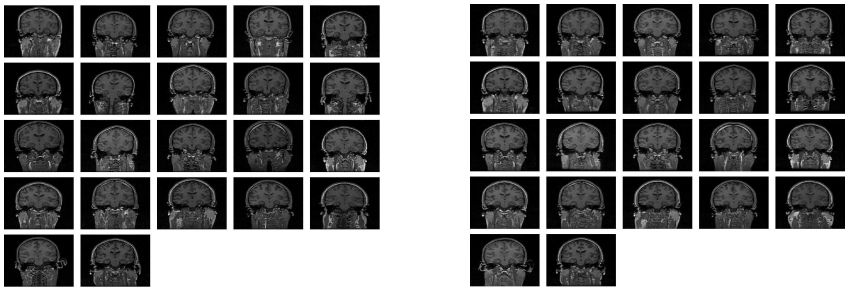
(a) Data set of 127 MRI volumes (b) Synthetic data set of 40 MRI volumes

**Fig. 1.** Orthogonal slices of the mean volume of the samples before and after alignment: (a)adult brain data set of 127 MRI volumes (b) synthetic data set of 40 MRI volumes.



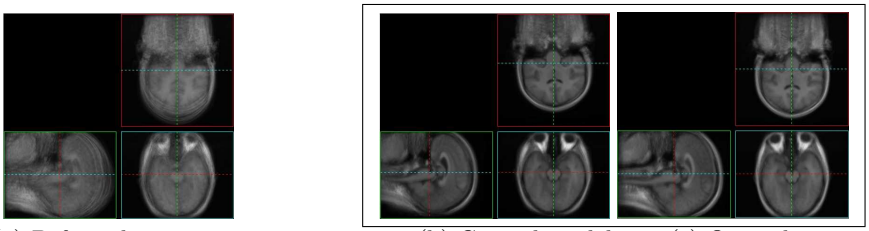
(a) Before alignment (b) After alignment

**Fig. 2.** Synthetic data set of 40 MRI volumes. Central slices of the input images (a) before and (b) after the population alignment.



(a) Before alignment (b) After alignment

**Fig. 3.** The adult brain data set of 22 MRI volumes used to make our atlas. Central slices of the input images (a) before and (b) after the population alignment.



(a) Before alignment (b) Control model (c) Our *atlas*

**Fig. 4.** 3D views of the mean volume created from the adult brain data population of 22 images: (a) before population alignment (b) the control model and (c) the model estimate of our algorithm.