Image-Driven Population Analysis Through Mixture Modeling

Mert R. Sabuncu*, *Member, IEEE*, Serdar K. Balci, Martha E. Shenton, and Polina Golland

***Abstract*—We present iCluster, a fast and efficient algorithm that clusters a set of images while co-registering them using a parameterized, nonlinear transformation model. The output of the algorithm is a small number of template images that represent different modes in a population. This is in contrast with traditional, hypothesis-driven computational anatomy approaches that assume a single template to construct an atlas. We derive the algorithm based on a generative model of an image population as a mixture of deformable template images. We validate and explore our method in four experiments. In the first experiment, we use synthetic data to explore the behavior of the algorithm and inform a design choice on parameter settings. In the second experiment, we demonstrate the utility of having multiple atlases for the application of localizing temporal lobe brain structures in a pool of subjects that contains healthy controls and schizophrenia patients. Next, we employ iCluster to partition a data set of 415 whole brain MR volumes of subjects aged 18 through 96 years into three anatomical subgroups. Our analysis suggests that these subgroups mainly correspond to age groups. The templates reveal significant structural differences across these age groups that confirm previous findings in aging research. In the final experiment, we run iCluster on a group of 15 patients with dementia and 15 age-matched healthy controls. The algorithm produces two modes, one of which contains dementia patients only. These results suggest that the algorithm can be used to discover subpopulations that correspond to interesting structural or functional "modes."**

***Index Terms*—Clustering, computational anatomy, image registration, population analysis, segmentation.**

Manuscript received December 24, 2008; revised March 03, 2009. First published March 24, 2009; current version published August 26, 2009. This work was supported by the Department of Veterans Affairs Merit Awards, National Alliance for Medical Image Analysis (NIH NIBIB NAMIC U54-EB005149), in part by the Neuroimaging Analysis Center (NIH CRR NAC P41-RR13218), in part by the Morphometry Biomedical Informatics Research Network (NIH NCRR mBIRN U24-RR021382), in part by the NIH NINDS R01-NS051826 grant, in part by National Institute of Mental Health Grant 5R01-MH050740-13, and in part by the National Science Foundation under CAREER Grant 0642971. *Asterisk indicates corresponding author.*

*M. R. Sabuncu is with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: msabuncu@csail.mit.edu).

S. K. Balci is with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: serdar@csail.mit.edu).

M. E. Shenton is with the Surgical Planning Laboratory, Harvard Medical School and Brigham and Womens Hospital, Boston, MA 02115 USA, with the Psychiatry Neuroimaging Laboratory, Department of Psychiatry, Brigham and Womens Hospital, Harvard Medical School, Boston, MA 02115, USA, and also with the Clinical Neuroscience Division, Laboratory of Neuroscience, VA Boston Healthcare System and Harvard Medical School, Brockton, MA 02301 USA.

P. Golland is with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: polina@csail.mit.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2009.2017942

I. Introduction

TODAY, computational anatomy studies are mainly hypothesis-driven, aiming to identify and characterize structural or functional differences between, for instance a group of patients with a specific disorder and control subjects. This approach is based on two premises: accurate clinical classification of subjects and spatial correspondence across the images. In practice, achieving either can be challenging. First, the complex spectrum of symptoms of neuro-degenerative disorders like schizophrenia and overlapping symptoms across different types of dementia, such as Alzheimer's disease, delirium and depression, make a diagnosis based on standardized clinical tests difficult [22]. Second, establishing across-subject correspondence in the images is a particularly hard problem constrained by the specifics of the application. A popular technique is to normalize all subjects into a standard space, such as the so-called Talairach space [47], by registering each image with a single, universal template image that represents an average brain [12]. However, the quality of such an alignment is limited by the accuracy with which the universal template represents the population in the study.

With the increasing availability of medical images, data-driven algorithms offer the ability to probe a population and potentially discover subgroups that may differ in unexpected ways. In this paper, we propose and demonstrate an efficient probabilistic clustering algorithm, called **iCluster**, that

1) computes a small number of templates that summarize a given population of images;
2) simultaneously co-registers all the images using a nonlinear transformation model;
3) assigns each input image to a template that best describes the image.

The templates are guaranteed to live in an affine-normalized space, i.e., they are spatially aligned with respect to an affine transformation model. A preliminary version of iCluster was published at the International Conference on Medical Image Computing and Computer Assisted Intervention [42]. This paper expands the conference paper with a more detailed theoretical development and more extensive experimental work.

In our experiments, we demonstrate that the templates computed by the proposed algorithm can be used for various purposes, including constructing multiple atlases for improved segmentation and discovering structural modes of a population. On a data set of 50 brain MR images with manual labels for several temporal lobe structures, we illustrate that the subpopulations computed by iCluster manifest significantly improved average label alignment compared to the clinical subpopulations and the whole population. This result suggests that a multi-

template strategy will yield improved segmentation accuracy in an atlas-based framework. In other experiments, we show that the modes of the population discovered by iCluster capture known structural differences and similarities. On a population of 415 brain magnetic resonance imaging (MRI) of subjects aged 18–96 years, the algorithm computed three unique templates that mainly comprised of young subjects (mean age 31), older middle aged subjects (mean age 69), and elderly subjects (mean age 79). In another setting, we demonstrate that the modes discovered by the algorithm reflect the two groups of subjects (with mild dementia and healthy) in the population. These results suggest that iCluster can be used to probe a population of images to discover important structural or functional "modes."

The remainder of the paper is organized as follows. Section II includes an overview of the literature on atlas construction and inter-subject registration. In Section III, we introduce the generative model and develop our algorithm. Section IV reports experimental results. Section V discusses the advantages and drawbacks of the proposed algorithm, while pointing to future directions of research. Section VI concludes with a summary of contributions.

II. Background and Prior Work

In medical imaging, the term *atlas* usually refers to a (probabilistic) model of a population of images, with the parameters learned from a training data set [14], [51]. In its simplest form, an atlas is a mean intensity image, which we call a template [6], [12], [53], [54]. Richer statistics, such as intensity variance or segmentation label counts, can also be included in the atlas model [19]. Atlases are used for various purposes including normalization of new subjects for structure and function localization, segmentation, or parcellation of certain structures of interest, and group analysis that aims to identify pathology-related changes or developmental trends.

Atlas construction requires a dense correspondence across subjects. Earlier techniques used a single image—either a standard template [12], or an arbitrary subject from the training data set [25]—to initially align images using a pairwise registration algorithm. Other methods focused on determining the least biased template from the training set [31], [37]. A single template approach faces substantial methodological challenges when presented with a heterogeneous population, such as patients and matched normal control subjects in clinical studies. To circumvent this, more recent approaches co-register the group of images simultaneously without computing a group template [46], [58]. Even though these algorithms remove the requirement of a single template, they do not attempt to model the heterogeneity in the population. Recent work [9] presented a method that automatically identified the modes of a population using a mean-shift algorithm. This approach solved pairwise registrations to compute each interimage distance, which slowed down the algorithm substantially. Furthermore, the multi-modality of the population was not modeled explicitly, making it difficult to extract a representation of the heterogeneous population. An alternative strategy to atlas-based segmentation is to use all training images as the atlas [27]. A new subject is registered with each training image and segmentation is based on a fusion of the manual labels in the training data. This approach is not suitable for anatomical variability studies, where a universal coordinate frame is necessary to identify and characterize group differences and study developmental and pathological trends.

There is a rich range of techniques used to characterize similarities and differences across subpopulations defined by attributes like gender, handedness and pathology. Volume-based [11], [39], [44], voxel-based [4], [15], and deformation-based [5] morphometry methods are commonly used to compare anatomical MRI scans of two or more groups of subjects. Other examples include statistical analysis of functional magnetic resonance imaging (fMRI), positron emission tomography (PET), and diffusion data to identify age and disease-related changes in the functional and structural organization of the brain [24], [33]. In these studies, intersubject correspondence is typically achieved via one of the image registration algorithms discussed above. When faced with a heterogeneous group of healthy and pathological brains, however, establishing intersubject correspondence is an ambiguous and more challenging problem due to dramatic structural changes associated with the pathology. For instance, defining a similarity measure when certain corresponding regions are missing or unclear, is not straightforward.

Probabilistic atlases are powerful tools used commonly for supervised segmentation [3], [13], [18], [55]. A probabilistic atlas can provide statistics about the frequency of a certain label at a particular location, and topological information like the frequency of two different labels neighboring each other at a particular location and with a certain orientation. Moreover, it can include information about the relationship between labels and image intensities. Given a new image, intensity models, such as a template image, are typically used for spatial normalization. Automatic segmentation is then formulated as an inference problem. Recent joint registration and segmentation frameworks [3], [38] integrate the two steps: spatial normalization is updated based on the current segmentation and vice versa. Most atlas-based segmentation approaches make a strong unimodal assumption on the intensity distribution either when building the atlas, or when segmenting the new image or at both stages. In other words, they assume a homogeneous population, where each subject can be modeled as a deformed and noisy version of a universal template. However, there is growing evidence that population-specific atlases can improve the quality of segmentation [48], [57]. This, we believe, highlights the limitations of a single-template atlas in segmentation applications and points toward a multi-template atlas strategy.

In this paper, we develop a probabilistic framework for joint registration of a set of images into a common coordinate frame, while clustering them into a small number of groups, each represented by a template image. We employ a mixture of Gaussians model and a maximum likelihood framework which we solve using the generalized expectation maximization (GEM) algorithm. A similar approach was independently developed in [1], which provides a rigorous analysis of the maximum a posteriori estimate of the deformable templates using a Gaussian kernel based deformation parametrization. In [1] the application of the framework was limited to 2-D images of handwritten digits. In contrast, we focus on high-resolution

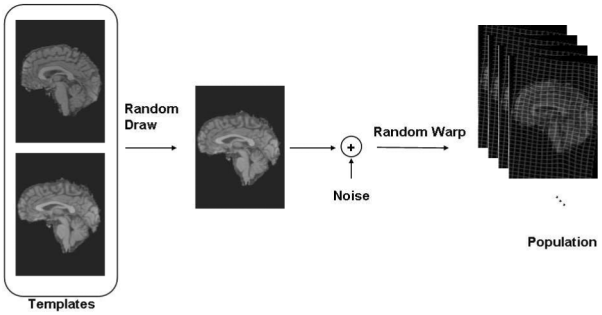


Fig. 1. Generative model that assumes two templates.

3-D medical data and employ a *B*-spline parametrization for the nonlinear transformation, as previously demonstrated in [41]. Furthermore, we present approximate solutions to the template estimation problem that yield fast algorithms applicable to large data sets. Our algorithm can also be viewed as an extension of the approach in [50], which solves the registration problem as an initial, separate step. Our framework leads to a fast, scalable, and flexible algorithm that removes the sensitivity of the resulting atlas coordinate frame to the selected target. Moreover, it provides a novel, data-driven way to probe the population for different modes. Analyzing the discovered subpopulations and their representative templates promises to advance our understanding of dominant structural or functional changes due to pathology or development.

III. Model and Algorithm

We assume that the input images $\{I_n\}_{n=1}^N$ are generated from a small number of templates $\{T_k\}_{k=1}^K$, where K is known and fixed. Later, we will propose a strategy to automatically determine K from the data. Thus, for each $n \in \{1,\ldots,N\}$, there exists $k \in \{1,\ldots,K\}$ such that

$$I_n(\vec{x}) = T_k\left(\Phi_n^{-1}(\vec{x})\right) + \epsilon_n\left(\Phi_n^{-1}(\vec{x})\right), \quad \forall \vec{x} \in \Omega \subset \mathbb{R}^3 \tag{1}$$

where $\Phi_n : \mathbb{R}^3 \mapsto \mathbb{R}^3$ is an admissible, invertible spatial warp, such as a parameterized nonlinear transformation, Φ_n^{-1} denotes its inverse, $\epsilon_n(\cdot)$ is a spatially independent, non-stationary Gaussian noise field with zero mean and standard deviation $\sigma(\cdot)$. The last term models imaging noise, and the independent Gaussian assumption is a commonly used model in the literature [18]. We model the noise parameters in the coordinate frame of the template. Fig. 1 illustrates this generative model for two templates.

Let $p_k(I_n; T_k, \sigma, \Phi_n)$ denote the conditional probability of the image I_n given that it is generated by the k'th template, and with the fixed model parameters. This can be computed from (1)

$$p_k(I_n; T_k, \sigma, \Phi_n) = \prod_{\vec{x}\in\Omega} \mathcal{N}\left(I_n(\vec{x}); T_k\left(\Phi_n^{-1}(\vec{x})\right) \sigma\left(\Phi_n^{-1}(\vec{x})\right)\right) \tag{2}$$

where $\mathcal{N}(\cdot;\mu,\sigma)$ is the Gaussian density with mean μ and standard deviation σ.

Let $\{\pi_k\}$ denote the prior probabilities of the templates. This distribution governs the initial random draw of templates shown in Fig. 1 and models the possibly unbalanced sizes of the clusters. Thus the parameters for the whole model include the templates $\{T_k\}$, template priors $\{\pi_k\}$ and standard deviation image $\sigma(\cdot)$. The spatial transformations $\{\Phi_n\}$ can be viewed as hidden random variables, drawn independently for each image from a prior distribution that favors smoother transformations, for instance. In this paper, however, for simplicity we will treat $\{\Phi_n\}$ as model parameters. We use $\theta = \{\{T_k\}, \{\pi_k\}, \sigma, \{\Phi_n\}\}$ to denote the pooled set of model parameters and spatial transformations. Marginalizing over all possible template indices, we obtain the probability of observing a particular image I_n

$$\begin{aligned} p(I_n;\theta) &= \sum_k \pi_k p_k(I_n; T_k, \sigma, \Phi_n) \\ &= \sum_k \pi_k \prod_{\vec{x}\in\Omega} \mathcal{N}\left(I_n(\vec{x}); T_k\left(\Phi_n^{-1}(\vec{x})\right), \sigma\left(\Phi_n^{-1}(\vec{x})\right)\right). \end{aligned} \tag{3}$$

A. Generalized EM for Atlas Construction

We formulate the problem of atlas construction as a maximum likelihood estimation

$$\theta^* = \operatorname*{argmax}_\theta L(\theta) = \operatorname*{argmax}_\theta \sum_n \log p(I_n;\theta) \tag{4}$$

where $L(\theta)$ denotes the log-likelihood of the entire image set evaluated for the parameter θ. We use a generalized expectation maximization (GEM) algorithm to solve (4). For a fixed $\theta_0 = \{\{T_{k0}\}, \{\pi_{k0}\}, \sigma_0, \{\Phi_{n0}\}\}$, using Jensen's inequality we form a lower bound for $L(\theta)$

$$\begin{aligned} L(\theta) &\geq Q(\theta;\theta_0) \\ &= \sum_n \sum_k q_k(I_n;\theta_0) \log \pi_k p_k(I_n; T_k, \sigma, \Phi_n) + c \end{aligned} \tag{5}$$

where c is a constant that does not depend on θ and $q_k(I_n;\theta_0)$ is the posterior probability that the image I_n was generated from the template T_k

$$q_k(I_n;\theta_0) = \frac{\pi_k p_k(I_n; T_{k0}, \sigma_0, \Phi_{n0})}{\sum_{k'} \pi_{k'} p_{k'}(I_n; T_{k'0}, \sigma_0, \Phi_{n0})}. \tag{6}$$

Note that $L(\theta_0) = Q(\theta_0;\theta_0)$. The GEM algorithm iteratively improves this lower bound. Let $\theta^{(i)}$ be the guess of θ at iteration i. Computing $Q(\theta;\theta^{(i)})$—or, equivalently $q_k(I_n;\theta^{(i)})$—is the E-step of iteration $i+1$. The M-step updates θ to increase $Q(\theta;\theta^{(i)})$. In our formulation, we use a coordinate ascent strategy in the M-step and divide it into two substeps: the T-step ("T" stands for template) where we compute the closed form expressions of the template parameters $\{\{T_k\}, \{\pi_k\}, \sigma\}$ that maximize $Q(\cdot;\theta^{(i)})$; and the R-step ("R" stands for registration) where we numerically solve for the transformation parameters $\{\Phi_n\}$. We will use $J(\Phi, \vec{x})$ to denote the Jacobian field of a transformation $\Phi(\vec{x})$ with respect to the spatial coordinates and $|J|$ will indicate the determinant of matrix J. Derivations for the T- and R-steps can be found in the Appendix. Here we summarize the algorithm.

- **E-step:** Given the model parameters from iteration i, the algorithm updates the posterior cluster probabilities:
  1) $\hat{q}_k(I_n;\theta^{(i)}) \propto \pi_k^{(i)} p_k(I_n; T_k^{(i)}, \sigma^{(i)}, \Phi_n^{(i)})$, where $p_k(\cdot)$ is defined in (2).

2) Normalize q_k to sum to 1

$$q_k\left(I_n;\theta^{(i)}\right)=\frac{\hat{q}_k\left(I_n;\theta^{(i)}\right)}{\sum_{k'}\hat{q}_{k'}\left(I_n;\theta^{(i)}\right)}. \tag{7}$$

These probabilities can be seen as "soft cluster memberships," where $q_k(I_n;\theta^{(i)})=1$ indicates a "hard membership" in cluster k.

- **T-step:** Given the posterior probability estimates $\{q_k(I_n;\theta^{(i)})\}$ and transformation parameters $\{\Phi_n^{(i)}\}$, the algorithm updates its estimates of the templates $\{T_k\}$, template priors $\{\pi_k\}$ and standard deviation image σ, for which we derive closed-form expressions

$$T_k^{(i+1)}(\vec{x})=\frac{\sum_n q_k\left(I_n;\theta^{(i)}\right)\left|J\left(\Phi_n^{(i)},\vec{x}\right)\right|\left(I_n\left(\Phi_n^{(i)}(\vec{x})\right)\right)}{\sum_n q_k\left(I_n;\theta^{(i)}\right)\left|J\left(\Phi_n^{(i)},\vec{x}\right)\right|} \tag{8}$$

$$\pi_k^{(i+1)}=\frac{1}{N}\sum_n q_k\left(I_n;\theta^{(i)}\right) \tag{9}$$

$$(\sigma^{(i+1)}(\vec{x})^2=\sum_{n,k}\frac{q_k\left(I_n;\theta^{(i)}\right)\left|J\left(\Phi_n^{(i)},\vec{x}\right)\right|\left(I_n\left(\Phi_n^{(i)}(\vec{x})\right)-T_k^{(i+1)}(\vec{x})\right)^2}{\sum_{n,k} q_k\left(I_n;\theta^{(i)}\right)\left|J\left(\Phi_n^{(i)},\vec{x}\right)\right|}. \tag{10}$$

- **R-step:** Given the new template parameters $\{T_k^{(i+1)}\},\{\pi_k^{(i+1)}\}$, standard deviation image $\sigma^{(i+1)}$, and memberships $\{q_k(\cdot;\theta^{(i)})\}$ the spatial transformations are updated

$$\Phi_n^{(i+1)}=\operatorname*{argmin}_{\Phi}\sum_{\vec{x}\in\Omega}|J(\Phi,\vec{x})|\times\frac{\left(I_n(\Phi(\vec{x}))-\bar{T}_n^{(i+1)}(\vec{x})\right)^2}{\sigma^{(i+1)}(\vec{x})^2} \tag{11}$$

$$=\operatorname*{argmin}_{\Phi}R_{\sigma^{(i+1)}}\left(I_n(\Phi),\bar{T}_n^{(i+1)}\right) \tag{12}$$

where $\bar{T}_n^{(i+1)}=\sum_k q_k(I_n;\theta^{(i)})T_k^{(i+1)}$ is the "effective template" (i.e., target image in registration) for image I_n at iteration $(i+1)$ and $R_\sigma(\cdot,\cdot)$ is the weighted sum of square differences (WSSD) objective function of the R-step. The effective template is a weighted average of the current templates and the weights are membership probabilities. A single, invertible transformation Φ_n is estimated for each image. Current membership probabilities determine which template the image should be aligning with.

We employ a *B*-spline transformation model (on an $8\times8\times8$ control point grid, unless specified otherwise) and a multiresolution strategy. In general, this transformation model does not guarantee invertibility. In practice, the algorithm checks for invertibility by monitoring the Jacobian terms and terminates when there is a Jacobian determinant value below a certain small positive threshold. Rather than solving the nonconvex optimization problem of (11), we perform a single Brent's method line search [10] based on gradient directions. The line search of each image is done in parallel, since the optimization for one image does not depend on other images. This strategy guarantees that the lower bound on the log-likelihood is improved, if not maximized, at each step; hence the name *Generalized* EM.

B. Initialization

The above GEM algorithm does not guarantee that the computed template images are in alignment. To introduce a notion of *common coordinate frame*, we use an initial affine normalization step that coregisters all images using a single dynamic mean image and an affine transformation model. This step is one of the popular coregistration algorithms used in practice. After affine normalization, the GEM algorithm starts with the E-step by computing membership probabilities according to (7). We initialize the template images as a random selection of K different input images, where K is the predetermined number of templates. In our experiments, we explore various values for K and only report results for the K values that produce robust results across multiple random initializations as discussed in Section IV-A. The template priors are initially assigned to be $1/K$, and the variance image is initialized to be the sample variance at each voxel after affine normalization. Each R-step is initialized with the transformation parameters from the previous iteration.

C. Gradient Re-Normalization

In group-wise registration, one needs to anchor the registration parameters to avoid global transformation drifts across subjects [8], [46], [58]. A natural common coordinate frame can be defined as the average of the population. This natural coordinate frame is computed implicitly by constraining the sum of all displacements across the subjects to be zero. We extend this strategy to the multi-template setting by constraining each point in the common coordinate frame to lie at the average location of corresponding points across the images *in each cluster*. To impose this constraint, we use the soft memberships $q_k(\cdot)$

$$\frac{\sum_n q_k\left(I_n;\theta^{(i)}\right)\Phi_n^{(i+1)}(\vec{x})}{\sum_n q_k\left(I_n;\theta^{(i)}\right)}=\vec{x},\quad\forall\vec{x}\in\Omega,\text{ and }\forall k. \tag{13}$$

Equivalently

$$\sum_n q_k\left(I_n;\theta^{(i)}\right)\Phi_n^{(i+1)}(\vec{x})=\sum_n q_k\left(I_n;\theta^{(i)}\right)\vec{x} \tag{14}$$

$\forall\vec{x}\in\Omega$, and $\forall k$. Summing both sides of (14) over k yields

$$\frac{1}{N}\sum_n\Phi_n^{(i+1)}(\vec{x})=\vec{x},\quad\forall\vec{x}\in\Omega \tag{15}$$

which is the anchoring constraint used by other group-wise registration methods [8], [46], [58].

In a gradient descent optimization strategy, one way of imposing the constraint of (15) is to re-normalize the gradients of the *R*-step objective function by subtracting the average gradient from all the individual image gradients. Let $\vec{g}_n^{(i+1)}=\nabla R_{\sigma^{(i+1)}}(I_n(\Phi),\bar{T}_n^{(i+1)})$ be a D dimensional row vector that

denotes the gradient of the *R*-step objective function with respect to the transformation parameters of the image I_n at iteration $i+1$. Then, before each update of the transformation parameters, one re-normalizes the gradients:

$$\vec{g}_n^{(i+1)} \leftarrow \vec{g}_n^{(i+1)} - \frac{1}{N}\sum_n \vec{g}_n^{(i+1)}. \tag{16}$$

In the multi-template setting, we extend this re-normalization to satisfy the constraint of (13). We stack all the gradient row vectors $\vec{g}_n^{(i+1)}$ to create an $N \times D$ matrix $G^{(i+1)}$ and all the membership probabilities $q_k(I_n;\theta^{(i)})$ to create an $N \times 1$ column vector $\vec{q}_k^{(i)}$ for each $k = 1,\ldots,K$. First, using the Gram-Schmidt process, we obtain at most K orthonormal vectors $\{\vec{u}_k^{(i)}\}$ from $\{\vec{q}_k^{(i)}\}_{k=1}^K$. Using this orthonormal basis, we re-normalize all the gradients as

$$\begin{aligned} G_j^{(i+1)} &\leftarrow \left[I_{N\times N} - \sum_k \vec{u}_k^{(i)}\left(\vec{u}_k^{(i)}\right)^T\right] G_j^{(i+1)} \\ &= G_j^{(i+1)} - \left[\sum_k \vec{u}_k^{(i)}\left(\vec{u}_k^{(i)}\right)^T\right] G_j^{(i+1)} \end{aligned} \tag{17}$$

where $I_{N\times N}$ denotes the $N \times N$ identity matrix, G_j denotes the jth column of G and $\vec{u}^T$ denotes the transpose of $\vec{u}$. After re-normalization each column of G is orthogonal to $\vec{q}_k^{(i)}$ for all k. In other words: $\left(G_j^{(i+1)}\right)^T \vec{q}_k^{(i)} = 0, \forall k = 1,\ldots,K$.

D. Determining the Optimal Number of Templates

Determining the optimal number of clusters is a classical problem in unsupervised machine learning, which unfortunately has no universal solution [35], [49]. The problem can be viewed as a specific case of model selection. In general, increasing the number of clusters provides a better fit to the observed data, yet this does not necessarily translate into improved generalization. A standard approach to controlling the generalization ability of the model is to penalize the model complexity. Bayesian information criterion (BIC) is a widely-used technique that attempts to achieve this balance [45]. In our setting, BIC (or equivalently minimum description length) can be formulated as minimizing the penalized negative log-likelihood

$$-2\log p(I_n;\theta(K)^*) + |\theta(K)|\log(N) \tag{18}$$

where $p(I_n;\theta(K)^*)$ is the maximum value of the likelihood in (3) for a fixed number of templates K and $|\theta(K)|$ is the total number of model parameters, which in our case is equal to $K + KV + V + ND$, where D is the number of transformation parameters and V is the number of voxels.

Alternatively, one can use the stability of the resulting model to quantitatively asses the structure in the clustered data, cf. [7]. In practice, we found it useful to measure the stability of the output against different random initializations. For example, we observed that beyond a particular input K, the computed clustering is significantly less consistent across runs with different initializations. We quantify this consistency using a relative measure defined *for each run* as

$$\frac{1}{N}\sum_n \sum_k q_k\left(I_n;\theta^{*(r)}(K)\right)\bar{q}_k(I_n;\theta(K)) \tag{19}$$

1. Affine normalization: Iteratively co-register all input images to a dynamic mean image with an affine transformation model.
2. Initialize the template images (with K random images), template priors (uniform) and variance image (intensity sample variance after Step 1)
3. Iterate until convergence:
   Sample a random subset of voxels. Using this set of samples:
   (i) E-step: Update membership probabilities using Equation (7).
   (ii) T-step: Update template images, priors and the variance image using Equations (8,9,10).
   (iii) R-step: Improve registration of images by performing a line search to decrease Equation (11) using the re-normalized gradients computed via Equation (17).

Fig. 2. iCluster: Pseudo-code.

where $q_k(I_n;\theta^{*(r)}(K))$ denotes the membership probabilities computed in run r and $\bar{q}_k(I_n;\theta(K))$ is the average membership probability over all remaining runs for a fixed input K. To handle the ambiguity in cluster indexing, we maximized (19) over all permutations of indexing of the templates in all runs. This procedure yields a relative consistency value for each run with a fixed input K. Based on the stability criterion, we propose to pick the highest value of K that yields a relatively high average consistency (e.g., the average over multiple runs exceeds 0.9).

We tested both BIC and the consistency criterion using synthetic data where ground truth was known. Our experiments, presented in Section IV-A, indicate that the consistency criterion yields an accurate prediction of the optimal number of templates.

E. Complexity

Each iteration of the algorithm has a computational complexity and memory requirement of $\mathcal{O}(NKV)$, where N is the number of input images, K is the number of templates and V is the number of voxels. We use multithreading in ITK [30] to implement a parallelized version of iCluster. Similar to [2], [58], we employ a stochastic subsampling strategy to speed up the algorithm. At each iteration, a random sample of less than 1% of the voxels was used to compute the soft memberships, templates, template priors, standard deviation image and to update transformation parameters. In practice, we run the numerical optimization of the *R*-step as a single line search for each image, where the search directions are the normalized gradients. The effect of stochastic subsampling is investigated using synthetic data in Section IV-A. Selecting a stopping criterion is not straightforward with the subsampling strategy, since a comparison of the objective function values across iterations is not possible. Instead, one can monitor the change in the parameters. In practice, the algorithm stops when the change in the class memberships and registration parameters falls below a predetermined threshold. Fig. 2 summarizes the iCluster algorithm.

IV. Experiments

We validate the algorithm and investigate its behavior in four different experiments. In the first experiment, we use synthetic data to inform a choice of parameter settings, including the amount of subsampling. The availability of ground truth allows us to quantify the quality of results objectively and perform comparisons across different settings of parameters. The second experiment demonstrates the use of iCluster for building a multi-

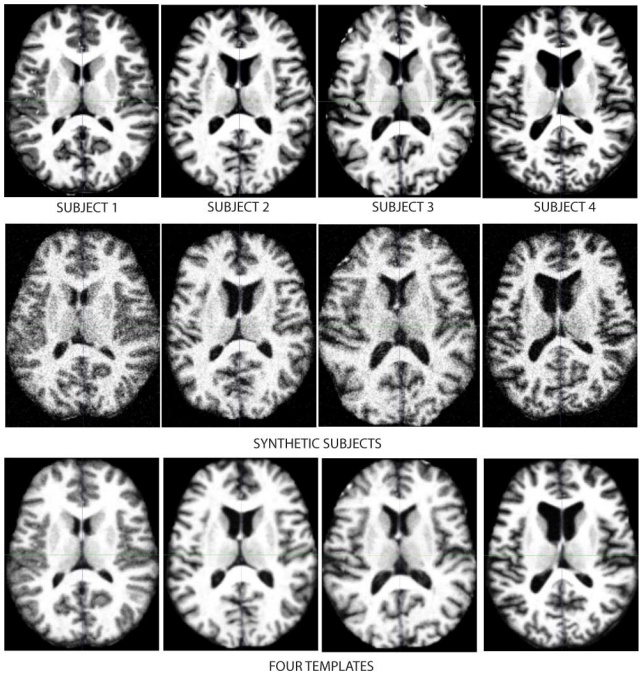


Fig. 3. Top row: Axial slices of the original subject MRIs used to synthesize data. Middle row: Axial slices of representative synthetic images. Bottom row: Axial slices of the four templates computed by iCluster with $K = 4$ and 0.5% sampling percentage.

template atlas for a segmentation application. In the third experiment, we employ iCluster to compute multiple templates from a large data set that contains 415 brain MRI volumes. Our results demonstrate that these templates correspond to different age groups. In the last experiment, we use our algorithm on a smaller population that contains patients with dementia and healthy subjects. The results indicate that the templates computed by the algorithm correspond to the two clinical groups. We find the correlation between the image-based clustering and demographic and clinical characteristics particularly intriguing, given that iCluster does not have access to this information when constructing the model of heterogeneity in the population.

A. Synthetic Experiments

In this experiment, we synthesized three data sets from four whole brain MR images (obtained from the Oasis repository [34], with an image resolution of $176 \times 208 \times 176$ voxels and voxel dimensions of 1 mm^3). The subjects were warped by applying random transformations parameterized with a $8 \times 8 \times 8$ *B*-spline model [41]. Each control point was displaced by an amount sampled uniformly from a 20 mm^3 box around its original location. Furthermore, the warped images were corrupted with i.i.d. zero mean Gaussian noise with a variance equal to 10% of the maximum intensity value. Axial slices of the original images and representative synthetic images are shown in Fig. 3. Table I summarizes the ground truth information for the synthetic data.

1) Effect of Stochastic Subsampling: First, we analyze the effect of stochastic subsampling on the quality of results. We ran iCluster on synthetic Data Set 3, with input $K = 4$. The four templates were initialized poorly as four different synthetic

TABLE I
SUMMARY OF GROUND TRUTH FOR THE SYNTHETIC DATA

	Templates	True K	# Subjects	Relative Cluster Size
Data Set 1	Subj. 1,4	2	15	0.4, 0.6
Data Set 2	Subj. 1,2,4	3	20	0.35, 0.5, 0.65
Data Set 3	Subj. 1-4	4	30	0.2, 0.3, 0.3, 0.2

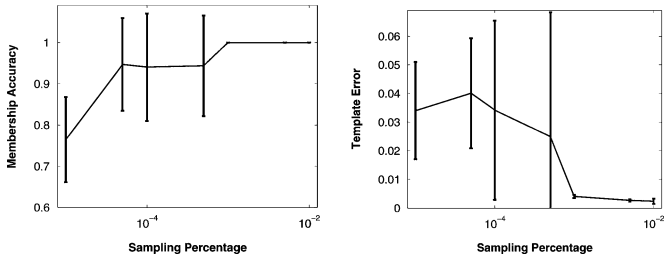


Fig. 4. Output quality as a function of sampling percentage, i.e., the ratio of the size of stochastic set of voxels used at each iteration to the total number of voxels. Error bars indicate standard deviation.

subjects that were all generated from the original subject 1. The quality of results was assessed using two measures: membership accuracy and error in the template images.

To define membership accuracy, we used the inner product between two membership probability matrices as a proxy for similarity. Formally, let $\{q_k(I_n;\theta)\}$ denote a set of output membership probabilities and $\{q_k^*(I_n)\}$ denote ground truth membership probabilities, with 1 corresponding to the template that generated the image and all remaining entries equal to zero. We define membership accuracy as

$$\frac{1}{N}\sum_n \sum_k q_k(I_n;\theta) q_k^*(I_n) \tag{20}$$

where N is the number of input images. To resolve the ambiguity in the cluster indices, we maximize (20) over all possible permutations of the ground truth template indices. We use this maximum value as a measurement of membership accuracy.

Let $T_k(\vec{x})$ denote the output template images and $T_k^*(\vec{x})$ denote the ground truth templates, i.e., original subject MRIs. We define the average template error as

$$\frac{1}{VK}\sum_k \sum_{\vec{x}\in\Omega} (T_k(\vec{x}) - T_k^*(\vec{x}))^2 \tag{21}$$

where V is the number of voxels in Ω, K is the number of templates and the template indexing is determined by maximizing (20) for output memberships.

Fig. 4 shows both the membership accuracy and template error values for a range of sampling percentages, where the sampling percentage is the ratio of the size of the stochastic set of voxels used at each iteration to the total number of voxels. For each parameter setting, we performed 10 runs of iCluster starting from the same poor initialization. Each run yielded a different output due to stochastic subsampling. For sampling percentage values larger than 0.1% membership accuracy was perfect and the template error reached its minimum for all ten runs. In practice, we chose 0.5% as the sampling percentage.

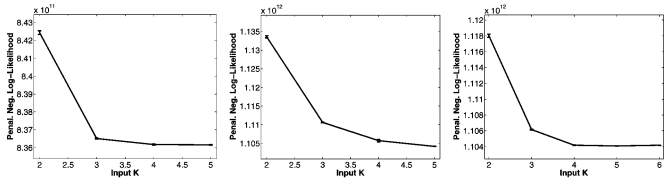


Fig. 5. BIC: Penalized negative log-likelihood values for a range of input K values. Error bars indicate standard error. (a) Synthetic Data 1. (b) Synthetic Data 2. (c) Synthetic Data 3.

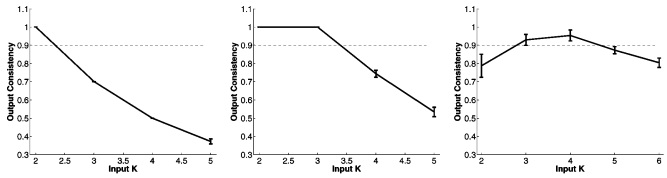


Fig. 6. Consistency Criterion: The consistency of output membership probabilities for a range of input K values. Error bars indicate standard error. (a) Synthetic Data 1. (b) Synthetic Data 2. (c) Synthetic Data 3.

This corresponds to using roughly 30 000 voxels at each iteration. The bottom row of Fig. 3 shows the four templates computed by iCluster with input $K = 4$ and 0.5% sampling percentage. The output templates were computed using (8) on the whole domain Ω with the estimated model parameters.

2) Determining the Optimal Number of Templates: Here, we compare two methods for automatically determining the optimal number of templates. We ran iCluster on the three synthetic data sets with a range of input K values. For each setting, we ran the algorithm ten times with different random initializations to get a collection of outputs. Using (18), we computed the negative penalized log-likelihood values for these outputs. Fig. 5 plots these values as a function of input K for the three data sets. BIC determines the optimal number of templates as the value of K that minimizes the penalized log-likelihood of the data under the estimated model. According to this criterion, data sets 1,2, and 3 have at least 4, 5, and 4 underlying templates, respectively. The optimal K for data sets 1 and 2 should have been 2 and 3, respectively.

Alternatively, we can look at the consistency of the resulting model to determine the optimal number of templates. We quantified the consistency of the model using the relative membership consistency measure defined in (19). The average relative membership consistency values for each input K are shown in Fig. 6. Based on the consistency criterion, we propose to select the highest value of K that yields a relatively high average consistency (e.g., the mean over multiple runs exceeds 0.9). According to this criterion, data sets 1, 2, and 3 have 2, 3, and 4 underlying templates, respectively, which agrees perfectly with the ground truth. In the remaining experiments, we used the consistency criterion to determine the optimal number of templates.

B. Segmentation Label Alignment

In atlas-based segmentation, one typically normalizes the new subject by registering the image with a template. Segmentation is then achieved by inferring labels based on the intensities of the new image and the training images that contain manual labels. The training data is usually employed to establish a prior for segmentation. To assess the quality of this prior, one can measure its agreement with the ground truth label of a new subject. In the following experiment, we measure this agreement by quantifying the alignment between one (new) subject and the remaining (training) subjects. In the case of multiple atlases, this requires an assignment of the new subject to one of the atlases. If these atlases are constructed through an image-based clustering strategy, as the one proposed in this paper, one can use the same framework to determine this assignment. This means fixing the template images, noise variance image and template priors in the iCluster algorithm. The assignment of the new subject can then be computed using the same GEM algorithm, which iterates over the E and R-steps.

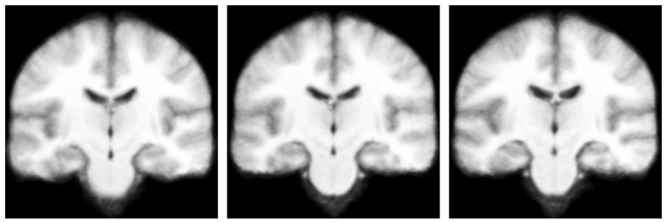


Fig. 7. Mean images for each clinical population after affine normalization. (a) Healthy controls. (b) Affective disorder. (c) Schizophrenia.

In this experiment, we used a data set of 50 whole brain MR brain images that contained 16 patients with first episode schizophrenia (SZ), 17 patients with first-episode affective disorder (AFF), and 17 age-matched healthy subjects (CON). The MRI volumes were obtained using a 1.5-T General Electric scanner (GE Medical Systems, Milwaukee, WI). The acquisition protocol was a coronal series of contiguous images. The imaging variables were as follows: $\mathrm{TR} = 35$ ms, $\mathrm{TE} = 5$ ms, one repetition, 45 nutation angle, 24-cm field-of-view, $\mathrm{NEX} = 1.0$ (number of excitations), matrix $= 256 \times 256$ (192 phase-encoding steps) $\times 124$. The voxel dimensions were $0.9375 \times 0.9375 \times 1.5$ mm. First episode patients are relatively free of chronicity-related confounds such as the long-term effects of medication, thus any structural differences between the three groups are subtle, local and difficult to identify in individual scans. Fig. 7 shows coronal slices of the affine-normalized mean images for each clinical population. A detailed description of the data and related findings are reported in [28].

For these images, we also had manual delineations of eight temporal lobe structures: the (left and right) superior temporal gyrus (STG), hippocampus (HIP), amygdala (AMY), and parahippocampal gyrus (PHG). Prior MRI studies of schizophrenic patients revealed structural brain abnormalities, with low volumes of gray matter in the left posterior superior temporal gyrus and in medial temporal lobe structures. However, the specificity to schizophrenia and the roles of chronic morbidity and neuroleptic treatment in these abnormalities remain under investigation [28], [29]. Accurate segmentation tools for temporal lobe structures is thus important for schizophrenia research. We used manual labels to explore label alignment across subjects under different groupings: on the *whole data*

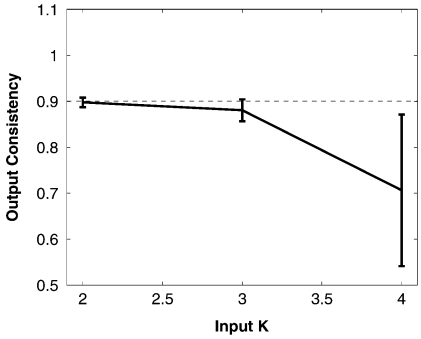


Fig. 8. Consistency Criterion for the schizophrenia data set: The consistency of output membership probabilities for input $K = 2, 3, 4$. Error bars indicate standard error.

TABLE II
CLINICAL COMPOSITION OF CLUSTERS FOR $K = 2$

Cluster	AFF	CON	SZ
1	11	9	10
2	6	8	6

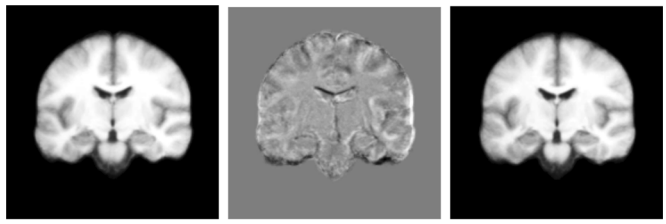
TABLE III
CLINICAL COMPOSITION OF CLUSTERS FOR $K = 3$

Cluster	AFF	CON	SZ
1	7	6	8
2	5	7	4
3	5	4	4

set, on the *clinical grouping*, and on the *image-based clustering* as determined by iCluster.

We ran iCluster on the 50 MR images for different values of input K. We emphasize that the algorithm did not have access to the clinical and manual label data. Fig. 8 shows the iCluster output membership consistency, as defined in Section III-D. We ran the algorithm ten times for each value of input K. Based on our proposed consistency criterion, we determine $K = 2$ as the optimal number of templates. However, to provide a comparison with the clinical grouping (where there are three groups: SZ, AFF, and CON), we present results for $K = 3$ as well. Tables II and III show the relationship between the clustering of the algorithm and the clinical diagnosis. We observe that the clustering computed by the algorithm demonstrates no correlation with the clinical diagnosis. This result confirms the difficulty of identifying structural differences between these first-episode patients and control subjects on an individual basis. Fig. 9 shows coronal views of the two templates discovered by iCluster and the difference image between these two. There are subtle structural differences between the two templates, especially around the cortical regions of the temporal lobes.

To measure the quality of alignment of a region of interest in two subjects, we employed two measures: 1) the Dice score which quantifies the overlap between the regions of interest in two subjects [55] and 2) the modified Haussdorff distance [56], which is defined as the average Euclidean distance (in millimeters) between a boundary point and the closest corresponding boundary point in the other subject. The Dice score ranges between 0 and 1, where 1 indicates a perfect overlap. The Haussdorff distance achieves zero at perfect alignment; higher values indicate worse alignment.



(a) (b) (c)

Fig. 9. Two Templates computed by iCluster. In the difference image, gray is zero, darker (lighter) values correspond to negative (positive) values. (a) Template 1. (b) Template 1 minus Template 2. (c) Template 2.

We compared average label alignments for three strategies.

1) **ALL**: All subjects were coregistered with a single dynamic average template. This was achieved using the iCluster algorithm with $K = 1$ and a $32 \times 32 \times 32$ *B*-spline grid. The average label alignment for each subject was then computed by averaging all pairwise measures of label alignment with the remaining subjects.
2) **CLIN**: Each clinical group was coregistered separately using iCluster with $K = 1$ and a $32 \times 32 \times 32$ *B*-spline grid. The average label alignment for each subject was then computed by averaging all pairwise measures of label alignment with the remaining subjects with the *same clinical diagnosis*.
3) **iC2** and **iC3**: We ran iCluster on all subjects with input $K = 2$ and 3, and a $32 \times 32 \times 32$ *B*-spline grid. For each input K value, we report label alignment results for the run that yielded the highest relative consistency value as defined in (19). The average label alignment for each subject was then computed by averaging all pairwise measures of label alignment with the remaining subjects in the *same cluster*.

Fig. 10 shows the average Dice scores and Hausdorff distances for the individual ROIs. These values were computed in the atlas space, where the manual labels were interpolated using the transformations obtained from the registrations and the nearest neighbor interpolator. We performed a paired permutation test comparison between the average label alignments of the three scenarios. The *p*-values were computed by assessing the average difference between two sets of paired measurements based on a histogram of differences obtained by randomly shuffling the order of pairings. The comparisons for the Haussdorff distances are presented in Table IV. Dice score comparisons yield similar results. In summary, iCluster with input $K = 2$ yields the best label alignment results, where 6 out of 8 ROIs were significantly better aligned (with $p < 0.05$) compared to the first two strategies of co-registering all subjects (ALL) and clinical groups separately (CLIN). This result provides further evidence for the usefulness of the proposed consistency criterion that determines the optimal number of templates. ALL and CLIN yield statistically improved label alignment for only one ROI: the right Superior Temporal Gyrus.

These results suggest that, on average, for most ROIs we achieve a better agreement between the ground truth labels and a prior obtained via iCluster, than a prior computed by co-registering all subjects or subjects within a clinical population.

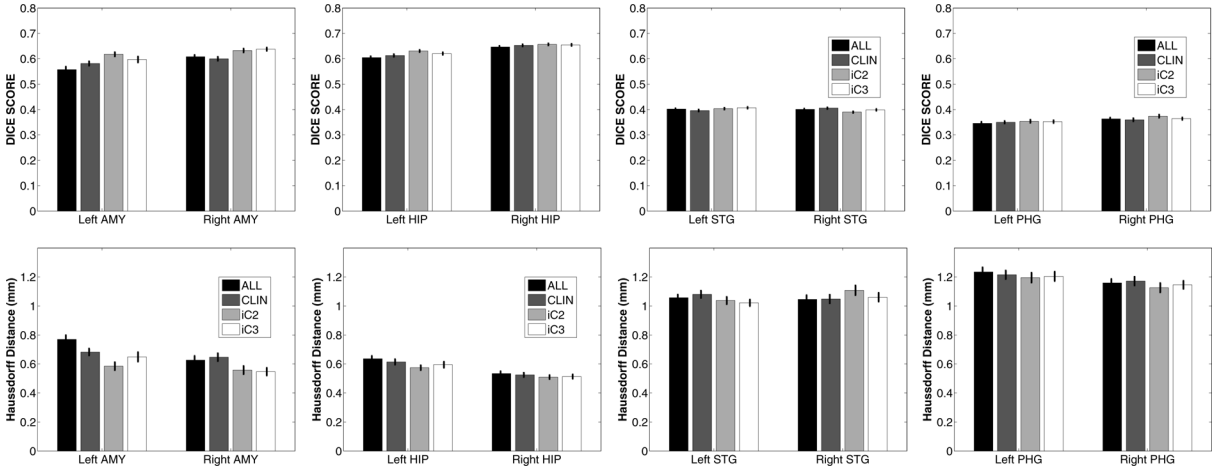


Fig. 10. Top row: Dice scores for each ROI. Bottom row: Haussdorff Distances in millimeters. Error bars indicate standard error.

TABLE IV
STATISTICAL COMPARISON OF AVERAGE LABEL ALIGNMENT. IMPROVEMENT: $+ + +p < 0.01, + + p < 0.05, +p < 0.1$. EQUIVALENT: $=$. IMPAIRMENT: $- - -p > 0.99$. L AND R DENOTE LEFT AND RIGHT, RESPECTIVELY

	l-AMY	r-AMY	l-HIP	r-HIP	l-STG	r-STG	l-PHG	r-PHG
iC2 vs. ALL	+++	+++	+++	++	+	- - -	++	+++
iC2 vs. CLIN	+++	+++	+++	+	+++	- - -	++	+++
iC3 vs. ALL	+++	+++	+++	++	+++	=	+	=
iC3 vs. CLIN	+++	+++	=	+	+++	=	=	+

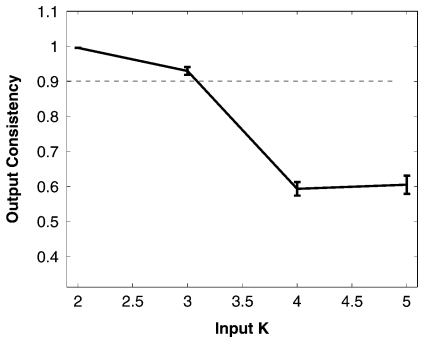


Fig. 11. Consistency Criterion for the Oasis data set: The consistency of output membership probabilities for a range of input K values. Error bars indicate standard error.

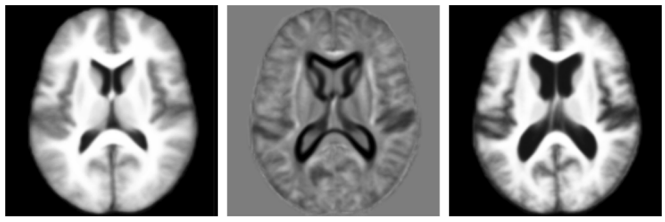


Fig. 12. Two templates of the OASIS data. In the difference image, gray is zero, darker (lighter) values correspond to negative (positive) values. (a) Template 1: Young. (b) Template 1 minus Template 2. (c) Template 2: Old.

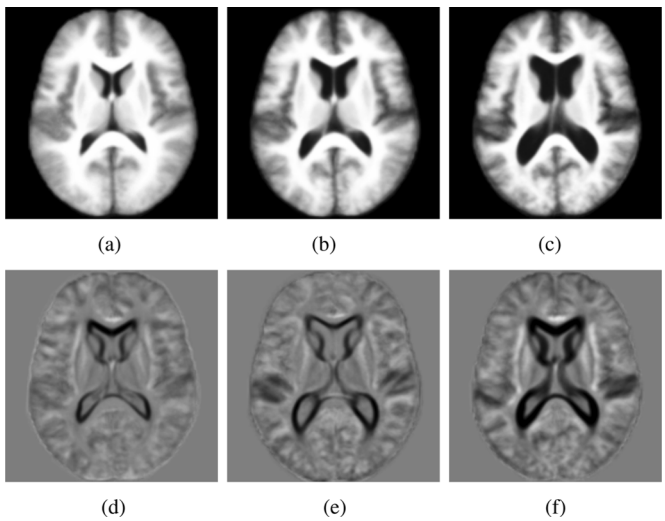
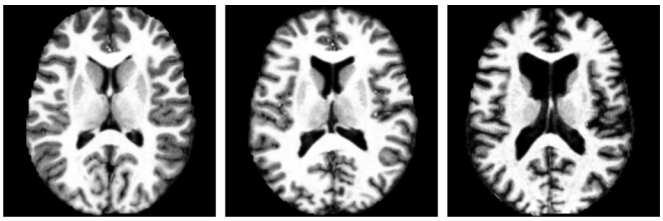


Fig. 13. Top Row: Three templates of the OASIS data. Bottom Row: Difference images. Gray is zero, darker (lighter) values correspond to negative (positive) values. (a) Template 1: Young. (b) Template 2: Older Middle Aged. (c) Template 3: Elderly. (d) Template 1 minus Template 2. (e) Template 2 minus Template 3. (f) Template 1 minus Template 3.

C. Age Groups in the OASIS Data Set

In this experiment, we used the OASIS data set [34] which consists of 415 preprocessed (skull stripped and gain-field corrected) brain MR images of subjects aged 18–96 years including individuals with early-stage Alzheimer's disease (AD). We ran iCluster on the whole data set while varying the number of templates K from 2 through 5. Each run took 4–8 h on a 16 processor PC with 128 GB RAM. Fig. 11 shows the output consistency against for different values of input K. For $K = 4$ and 5 the consistency values are significantly smaller than 0.9. We, therefore, report our results for $K = 2$ and $K = 3$. Figs. 12 and 13 show the two and three robust templates computed with $K = 2$ and $K = 3$, respectively. Fig. 14 shows typical individual subjects (in their native coordinates) corresponding to each cluster computed with $K = 3$. These subjects were chosen based on age, gender, and clinical condition, not image similarity. Fig. 15 shows the age distributions determined via Parzen



(a) (b) (c)

Fig. 14. Typical Subjects: (a) Group 1: 24-year-old, healthy female. (b) Group 2: 52-year-old, healthy female. (c) Group 3: 76-year-old male with very mild dementia and probable AD.

window estimator based on a Gaussian kernel with a standard deviation of four years.

It is easy to see that each template corresponds to a unique age group: For $K = 2$, we identify a group of 268 *young* subjects (aged 39.1 ± 19.9 years) and a group of 147 *elderly* subjects (aged 77.8 ± 9.3 years). For $K = 3$ the algorithm detected 201 *young* subjects (Group 1, aged 31.2 ± 14.5 years), an older *middle* aged group of 127 subjects (Group 2, aged 68.9 ± 13.6 years) and *elderly* 87 subjects (Group 3, aged 79.6 ± 7.5 years). Fig. 15(b) illustrates the intersection between the middle aged distribution of $K = 3$ and the distributions of $K = 2$. This plot reveals that the middle aged group for $K = 3$ consists of two subpopulations: 1) a younger group of subjects that are in the young group for $K = 2$ and 2) an older age group in the elderly for $K = 2$. These results suggest that the dominant structural modes in this large population are mainly due to aging. Analyzing the decomposition of the whole age distribution [shown in black in Fig. 15(b)] indicates that iCluster does not simply find the three major age modes. Specifically, the small middle peak around 50 years is robustly included with the younger population in both $K = 2$ and $K = 3$. With three modes, the algorithm identifies an older middle aged group (Group 2) that has a significant overlap in age with the elderly group (Group 3).

We further analyzed the clinical dementia rating (CDR) [36] data to explore the differences across the image-based clusters. Table V summarizes the results. Group 1 [Fig. 13(a)] has almost no subjects with positive CDR (an indication of probable Alzheimer's), whereas Group 2 [Fig. 13(b)] consists of 35% patients diagnosed with probable Alzheimer's disease (AD) (i.e., has a CDR score of greater than zero), and 65% subjects with no dementia. Group 3 [Fig. 13(c)] includes 69% patients with probable AD and 31% healthy subjects with zero CDR. The difference between the patient percentage in each group is statistically significant at $p < 10^{-4}$ as determined by a permutation test. This result indicates that the old-middle aged group computed by iCluster contains a majority of healthy individuals, whereas the elderly group is dominated by probable AD patients.

An important question at this point is to what extent these dementia profiles are correlated with the age data of the individuals, since it is known that the rate of incidence of dementia increases with aging [21]. Moreover, we would like to explore the influence of gender on these structural modes. One important point to note is that approximately half of the subjects over 60 years old (100 subjects) were clinically diagnosed with dementia, as summarized in Table VI. Examining this table reveals a difference between the two genders: healthy females without dementia are more likely to belong to Group 2 [Fig. 13(b)]. On the other hand, males with positive CDR (i.e., with dementia) are more likely to belong to Group 3 [Fig. 13(c)]. For the other two groups, i.e., males without dementia and females with dementia, there is no obvious relationship that these tables reveal.

To get a better insight into the characteristics of the discovered structural modes, we performed a multinomial logistic regression on the iCluster group memberships using age, gender and clinical data[1] as regressors. Table VII reports the regression coefficients, assuming Group 2 to be the reference category. If we convert the estimated probabilities to group assignments, the total model achieves around 75% training accuracy and a likelihood ratio test estimates the significance of the full, fitted model at $p < 0.01$. The significance of each coefficient was determined with a Wald test [17]. These results suggest that the most significant factor that determines group assignment is age: with each year, the odds of a subject being assigned to the next, older group increases by approximately $0.1 (\approx \exp(0.1) - 1)$. Groups 2 and 3 are also differentiated by the clinical score and gender (with less significance). One point decrease in the MMSE score increases the odds of a subject belonging to Group 3, rather than Group 2, by $0.1 (\approx \exp(0.1) - 1)$. A female's odds of belonging to Group 2 versus Group 3 is roughly 2.5-fold $(\exp(0.94))$ higher than a male's.

These results confirm that aging and dementia are both significant factors that influence major structural changes in the brain. Moreover, our results indicate that these factors may have different effects for the two genders. These findings demonstrate a qualitative similarity with the ones reported in [20], where aging and dementia are shown to correlate with brain atrophy in a similar manner. Furthermore, [20] reports that these effects have a tendency to be different in the two genders: males tend to demonstrate a higher rate of atrophy. The gender difference, however, does not reach statistical significance in the analysis of [20] and remains under debate in the literature [23], [32].

D. Patients With Dementia

In the fourth experiment, we used a set of 30 subjects (aged between 65 and 84 years) from the OASIS data set. Fifteen of these had a positive CDR, i.e., were diagnosed with very mild to mild dementia and probable AD (aged 74.7 ± 5.5 years, with education level of 3 ± 1.2), while the other 15 individuals were controls (aged 74.6 ± 5.4 years, with education level of 3.1 ± 1.5) with no sign of clinical dementia at the time of scanning. Fig. 16 shows the consistency of iCluster outputs over a range of input *K* values. For $K > 2$, we observe that the membership consistency is less than 0.9, thus we report results for $K = 2$: Group 1 [Fig. 17(a)] consists of 25 subjects, 15 of which were CDR zero. Group 2 [its template shown in Fig. 17(c)] consists of five subjects, all of which have *dementia*.

We performed a multinomial logistic regression on the iCluster assignments using age, education data (1: less than high school, 2: high school, 3: some college, 4: college graduate, 5: beyond college), clinical score and gender data as regressors. Only the clinical score demonstrated significant

[1]Mini-Mental State Exam scores [40] that ranged from 14 (poor mental health) to 30 (good mental health).

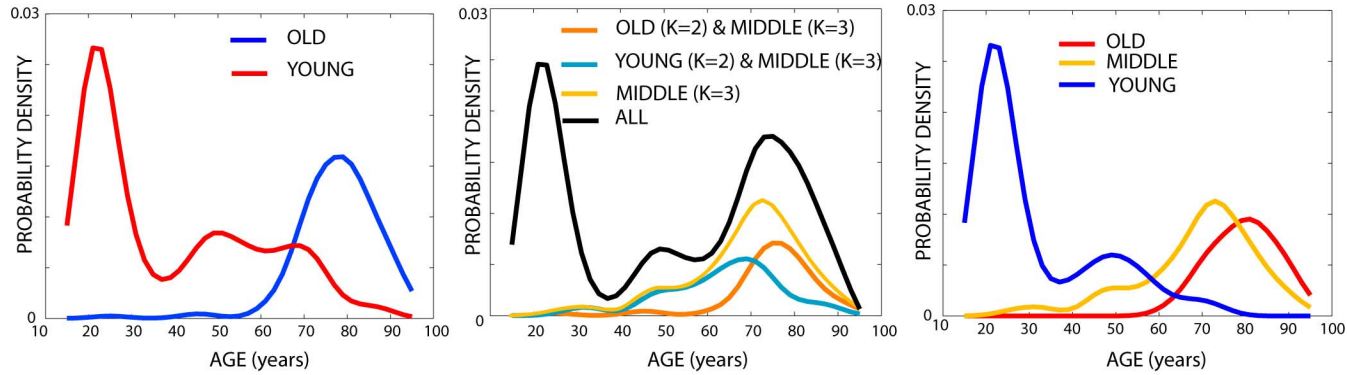


Fig. 15. Age distributions of the OASIS data. (a) Age distributions for $K = 2$, (b) the relationship between the ages of subjects in clusters identified for $K = 2$ and for $K = 3$, (c) Age distributions for $K = 3$. (a) $K = 2$, (b) from $K = 2$ to $K = 3$, (c) $K = 3$.

TABLE V
NUMBER (PERCENTAGE) OF SUBJECTS WITH RESPECT TO THEIR GENDER AND CLINICAL DEMENTIA SCORE IN EACH GROUP COMPUTED BY ICLUSTER WITH $K = 3$

	Positive CDR		Zero CDR	
	Female	Male	Female	Male
Group 1	1 (0.2)	1 (0.2)	119 (28.7)	80 (19.3)
Group 2	28 (6.8)	16 (3.9)	58 (14.0)	25 (6.0)
Group 3	30 (7.2)	24 (5.8)	19 (4.6)	14 (3.4)

TABLE VI
NUMBER (PERCENTAGE) OF SUBJECTS AGED 60 AND OLDER WITH RESPECT TO THEIR GENDER AND CLINICAL DEMENTIA SCORE DATA IN EACH GROUP COMPUTED BY ICLUSTER WITH $K = 3$

	Positive CDR		Zero CDR	
	Female	Male	Female	Male
Group 1	1 (0.5)	1 (0.5)	7 (3.5)	2 (1.0)
Group 2	28 (14.1)	16 (8.1)	46 (23.2)	10 (5.1)
Group 3	30 (15.2)	24 (12.1)	19 (9.6)	14 (7.1)

relevance to differentiate the two groups (see Table VIII): the first group's average MMSE score was 25.2 ± 5.1, whereas group two's score was 19.8 ± 2.9.

The fact that Group 2 comprised of dementia patients with significantly low MMSE scores is intriguing. Yet, the more interesting question is, what is special about the ten dementia patients assigned to Group 1? This clustering suggests that their anatomies are more similar to healthy subjects in the same age group. Clinical and demographic attributes of the patients in the two groups are virtually identical: 1) age: 74.4 ± 4.9 versus 75.4 ± 7.2, 2) MMSE score: 19.8 ± 3.9 versus 19.8 ± 2.9, and 3) education level: 3 ± 1.2 versus 3 ± 1.4. Thus, based on the data we have, this question remains open and requires further investigation.

V. DISCUSSION

Our experiments demonstrate the use of iCluster in multiple settings. The synthetic experiments served to asses the effect of stochastic subsampling on the quality of results and informed the design of the method that automatically determines the optimal number of templates. In the second experiment presented in Section IV-B, we show that, using the proposed clustering strategy, one can compute a multi-template atlas for a segmentation application. Based on growing evidence that population-specific atlases yield more accurate segmentation, we can employ iCluster to discover coherent subpopulations in a large population of images and construct separate atlases for each subpopulation. Our experiments suggest that a multi-template atlas can improve segmentation quality. The proposed approach promises significantly better segmentation than a disease-specific atlas, especially in the case of spectrum diseases such as schizophrenia.

In another setting, we demonstrate the utility of an image-driven approach for computational anatomy. This is in contrast with today's popular techniques that rely on a clinical or demographic classification of the subjects. Our experiments show that iCluster can robustly identify structural modes in a population that are mainly determined by age and dementia. This type of analysis promises to provide insight into the major factors that drive structural change and, more importantly, characterize subtypes of a particular disorder.

In our experiments, enlarged ventricles are immediately obvious in the older and dementia templates when compared to the younger and healthy populations, respectively. Moreover, cortical thinning and anterior white matter changes are visible in the difference images shown in Fig. 13. These types of structural changes due to aging and dementia have been well documented in the literature [16], [26], [43]. Further analysis is required to understand the structural differences between the discovered modes. The intermediate group (the older middle aged in the first experiment) and the mixture group in the dementia experiment can provide interesting insights into structural changes due to aging and dementia.

With a single template, i.e., input $K = 1$, iCluster can be seen as an efficient unbiased template estimation algorithm, similar to the ones proposed in [14], [31], [58]. Yet, the main point of this paper is that a single template is not sufficient to summarize the variability in a large and heterogenous population of images. To that extent, iCluster is similar to the recent works on atlas stratification [9] and deformable templates [1]. In the atlas stratification framework of [9], the authors propose to use an off-the-shelf clustering algorithm on images to identify underlying homogeneous subpopulations. The framework does not explicitly model anatomical heterogeneity and yields a computationally inefficient algorithm, where one needs to perform $\mathcal{O}(N^2)$ pairwise registration instances to analyze N input images. The generative model we developed in this paper is similar

TABLE VII
LOGISTIC REGRESSION COEFFICIENTS ON ICLUSTER MEMBERSHIPS COMPUTED FOR THE WHOLE OASIS DATASET AND $K = 3$

	Age (years)	Clinical State (MMSE)	Gender (Male: 0, Female: 1)
Group 1 vs. Group 2	-0.12^{***}	0.34	0.84
Group 3 vs. Group 2	0.11^{***}	-0.10^{**}	-0.94^{*}

TABLE VIII
LOGISTIC REGRESSION COEFFICIENTS ON ICLUSTER MEMBERSHIPS COMPUTER FOR THE 30 SUBJECT DEMENTIA DATASET AND $K = 2$

	Age (years)	Clinical State (MMSE)	Gender (Male: 0, Female: 1)	Education
Group 1 vs. Group 2	-0.03	-0.36^{*}	-2.74	0.22

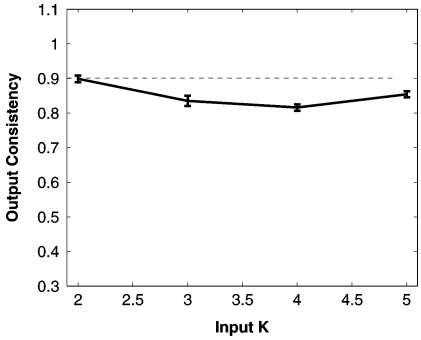
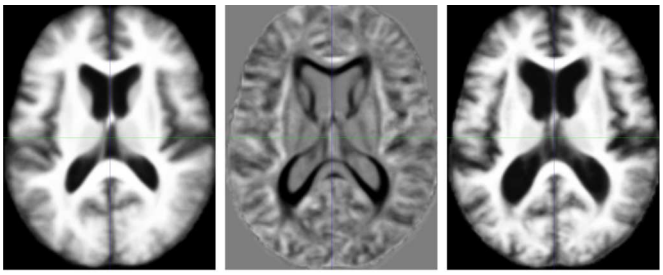


Fig. 16. Consistency criterion for the 30 subject dementia data set. The consistency of output membership probabilities for a range of input K values. Error bars indicate standard error.



(a) (b) (c)

Fig. 17. Two templates and their difference image for the 30 subject dementia data set. (a) Mostly Healthy. (b) Difference Image. (c) Dementia Patients.

to the deformable templates model of [1]. Yet, in contrast with [1], our main focus is to propose a computationally efficient algorithm that can be employed on large collections of high resolution medical image data. Most importantly, however, we include a concrete demonstration of how an image-clustering approach can be used to construct multiple segmentation atlases and study the effects of clinical and demographic factors on neuroanatomy.

The image-based clustering approach can also be extended to descriptors of anatomical shape, such as volume [20] or surface-based representations [52]. Various shape descriptors have been used to study the effects of disease progression and aging on anatomy. Based on similarity measures defined for these different descriptors, one can potentially derive different shape clustering algorithms. One such algorithm was proposed in [50]. The main drawback of such a shape-based approach is the need for accurate segmentations, which limits the amount of data such a strategy can be applied to. An image-based clustering approach, on the other hand, has the advantage that it can be used with large collections of raw images. Furthermore, image-based clustering can potentially reveal modes in a population that differ in unexpected regions.

We view iCluster as a first step towards a more comprehensive image-driven population analysis framework. The current algorithm suffers from several limitations. Notably, the simple additive Gaussian noise model cannot handle significant intensity variations across images. Thus, the current algorithm can only be used with intensity corrected (e.g., histogram matched, bias field corrected) images of the same modality. Moreover, the algorithm constructs clusters based on a similarity measure computed over whole images. This makes the method less sensitive to subtle and local differences across groups of images. One solution is to use a similarity measure computed over a region of interest in the E-step of iCluster. In the following, we summarize the possible directions one can explore to extend iCluster to a broader set of problems.

1) Use an entropy-based similarity measure that is insensitive to intensity variations to compute memberships in the E-step and perform co-registration in the R-step.
2) Compute memberships within a region of interest or based on a different type of information, e.g., connectivity from diffusion data.
3) Use more sophisticated models of deformation, e.g., diffeomorphisms. Moreover, one can integrate a more sophisticated prior on the spatial transformations. Hence, the memberships will be a function of both a similarity measure based on image intensities and the deformation cost.
4) Rather than using an additive noise model on intensities, one could explicitly model the variance in warps which would lead to a clustering strategy based on deformations.

VI. CONCLUSION

We presented a fast and efficient image clustering algorithm for co-registering a group of images, computing multiple templates that represent different modes in the population, and determining template assignments. We demonstrated our algorithm in several experiments, which illustrated a multi-template atlas strategy for accurate image segmentation and revealed age and disease-related modes in a population. Our results confirm previous findings and lead to interesting insights that suggest future research directions in computational anatomy.

APPENDIX I

In this Appendix, we provide derivations for the update equations of the T- and R-steps of the iCluster algorithm presented in Section III-A.

A. T-Step

Given the posterior probability estimates $\{q_k(I_n;\theta^{(i)})\}$ and fixing the spatial transformations $\{\Phi_n^{(i)}\}$ from the previous iteration, the template images $\{T_k\}$, template priors $\{\pi_k\}$, and standard deviation image σ are updated to maximize the lower bound $Q(\theta,\theta^{(i)})$ of (5)

$$\left\{\left\{T_k^{(i+1)}\right\},\left\{\pi_k^{(i+1)}\right\},\sigma^{(i+1)}\right\} = \underset{\{T_k\},\{\pi_k\},\sigma}{\operatorname{argmax}} \sum_n \sum_k q_k\left(I_n;\theta^{(i)}\right)\log \pi_k p_k\left(I_n;T_k,\sigma,\Phi_n^{(i)}\right) \tag{22,23}$$

such that $\sum_k \pi_k = 1$.

In (23) all the template priors $\{\pi_k^{(i+1)}\}$ can be optimized independently. We introduce a Lagrange multiplier λ for the constraint

$$\left\{\pi_k^{(i+1)}\right\} = \underset{\{\pi_k\}}{\operatorname{argmax}} \sum_n \sum_k q_k\left(I_n;\theta^{(i)}\right)\log \pi_k + \lambda\left(1-\sum_{k'}\pi_{k'}\right)+\text{const} \tag{24}$$

differentiate (23) with respect to π_k and set the derivative to zero, obtaining

$$\pi_k^{(i+1)} = \frac{1}{\lambda^*}\sum_n q_k\left(I_n;\theta^{(i)}\right) \tag{25}$$

where $\lambda^* = \sum_{k'}\pi_{k'}^{(i+1)} = \sum_{k',n} q_{k'}(I_n;\theta^{(i)}) = N$.

We recall that

$$\begin{aligned}
&\log p_k(I_n;T_k,\sigma,\Phi_n)\\
&\quad= -\left(\sum_{\vec{x}\in\Omega}\frac{\left(I_n(\vec{x})-T_k\left(\Phi_n^{-1}(\vec{x})\right)\right)^2}{2\sigma\left(\Phi_n^{-1}(\vec{x})\right)^2} + \log\sigma\left(\Phi_n^{-1}(\vec{x})\right)\right)+\text{const} &(26)\\
&\quad\approx -\int_{\Omega_c}\left(\frac{\left(I_n(\vec{x})-T_k\left(\Phi_n^{-1}(\vec{x})\right)\right)^2}{2\sigma\left(\Phi_n^{-1}(\vec{x})\right)^2} + \log\sigma\left(\Phi_n^{-1}(\vec{x})\right)\right)d\vec{x}+\text{const} &(27)\\
&\quad= -\int_{\Omega_c}\left(\frac{(I_n(\Phi_n(\vec{y}))-T_k(\vec{y}))^2}{2\sigma(\vec{y})^2} + \log\sigma(\vec{y})\right)|J(\Phi_n,\vec{y})|d\vec{y}+\text{const} &(28)\\
&\quad\approx \left(-\sum_{\vec{x}\in\Omega}\left(\frac{(I_n(\Phi_n(\vec{x}))-T_k(\vec{x}))^2}{2\sigma(\vec{x})^2} + \log\sigma(\vec{x})\right)|J(\Phi_n,\vec{x})|\right)+\text{const} &(29)
\end{aligned}$$

where $|\cdot|$ denotes matrix determinant, $J(\Phi,\vec{x})$ is the Jacobian matrix of Φ that contains the partial derivatives of the warp field with respect to the coordinates and Ω_c is a continuous and compact subset of $\mathbb{R}^3$ that covers the discrete set Ω. Equations (27)–(29) assume a suitable interpolator for making $I, \{T_k\}$ and σ spatially continuous. Equation (28) assumes the boundary condition $\Phi_n(\partial\Omega_c) = \partial\Omega_c$ for all n, where $\partial\Omega_c$ is the boundary of Ω_c and uses a change of variables with $y \triangleq \Phi_n^{-1}(\vec{x})$.

Substituting (29) into (23), we obtain

$$T_k^{(i+1)} = \underset{T_k}{\operatorname{argmin}} \sum_n\sum_{\vec{x}\in\Omega} q_k\left(I_n;\theta^{(i)}\right)\left|J\left(\Phi_n^{(i)},\vec{x}\right)\right| \times\frac{\left(I_n\left(\Phi_n^{(i)}(\vec{x})\right)-T_k(\vec{x})\right)^2}{2\sigma(\vec{x})^2}. \tag{30}$$

Differentiating the objective function in (30) with respect to $T_k(\vec{x})$ and setting the derivative to zero yields

$$T_k^{(i+1)}(\vec{x}) = \frac{\sum_n q_k\left(I_n;\theta^{(i)}\right) I_n\left(\Phi_n^{(i)}(\vec{x})\right)|J(\Phi_n^{(i)},\vec{x})|}{\sum_n q_k\left(I_n;\theta^{(i)}\right)\left|J\left(\Phi_n^{(i)},\vec{x}\right)\right|} \tag{31}$$

which is independent of $\sigma(\vec{x})$.

To determine $\sigma(\vec{x})$, we substitute (29) into (23) and obtain

$$\begin{aligned}
\sigma^{(i+1)} &= \underset{\sigma}{\operatorname{argmin}}\sum_n\sum_k q_k\left(I_n;\theta^{(i)}\right)\\
&\quad\times\sum_{\vec{x}\in\Omega}\left|J\left(\Phi_n^{(i)},\vec{x}\right)\right|\\
&\quad\times\left(\frac{\left(I_n\left(\Phi_n^{(i)}(\vec{x})\right)-T_k(\vec{x})\right)^2}{2\sigma(\vec{x})^2}+\log\sigma(\vec{x})\right).
\end{aligned}\tag{32}$$

Differentiating the objective function of (32) with respect to $\sigma(\vec{x})$ and setting the derivative to zero yields

$$\sigma^{(i+1)}(\vec{x})^2 = \frac{\sum_{n,k} q_k\left(I_n;\theta^{(i)}\right)\left|J\left(\Phi_n^{(i)},\vec{x}\right)\right|\left(I_n\left(\Phi_n^{(i)}(\vec{x})\right)-T_k^{(i+1)}(\vec{x})\right)^2}{\sum_{n,k} q_k\left(I_n;\theta^{(i)}\right)\left|J\left(\Phi_n^{(i)},\vec{x}\right)\right|}. \tag{33}$$

B. R-Step

Fixing the model parameters computed in the previous *T*-step, the *R*-step updates the transformations $\{\Phi_n\}$ to improve the lower bound $Q(\theta,\theta^{(i)})$ of (5). Substituting (29) into (23) and focusing on the terms that depend on Φ_n yields

$$\begin{aligned}
\Phi_n^{(i+1)} &= \underset{\Phi}{\operatorname{argmin}}\sum_k q_k\left(I_n;\theta^{(i)}\right)\sum_{\vec{x}\in\Omega}|J(\Phi,\vec{x})|\\
&\quad\times\frac{\left(I_n(\Phi(\vec{x}))-T_k^{(i+1)}(\vec{x})\right)^2}{\sigma^{(i+1)}(\vec{x})^2}
\end{aligned}\tag{34}$$

$$= \underset{\Phi}{\operatorname{argmin}} \sum_{\vec{x}\in\Omega} \frac{1}{\sigma^{(i+1)}(\vec{x})^2} |J(\Phi,\vec{x})| \times \left(I_n(\Phi(\vec{x}))^2 - 2I_n(\Phi(\vec{x})) \sum_k q_k \left(I_n;\theta^{(i)}\right) T_k^{(i+1)}(\vec{x}) \right) \tag{35}$$

$$= \underset{\Phi}{\operatorname{argmin}} \sum_{\vec{x}\in\Omega} |J(\Phi,\vec{x})| \times \frac{\left(I_n(\Phi(\vec{x})) - \sum_k q_k \left(I_n;\theta^{(i)}\right) T_k^{(i+1)}(\vec{x}) \right)^2}{\sigma^{(i+1)}(\vec{x})^2} \tag{36}$$

where in (35) and (36) we dropped and added terms that do not depend on Φ.

ACKNOWLEDGMENT

The authors would like to thank B. Fischl, K. Van Leemput, B. T. Thomas Yeo, and the anonymous reviewers for their helpful feedback. The authors would also like to thank to Dr. R. Buckner for making the OASIS dataset available.

REFERENCES

[1] A. Allassonniere, Y. Amit, and A. Trouve, "Towards a coherent statistical framework for dense deformable template estimation," *J. R. Stat. Soc.*, vol. 69, pp. 3–29, 2007.

[2] P. Viola and W. M. Wells, "Alignment by maximization of mutual information," *Int. J. Comput. Vis.*, vol. 24, no. 2, pp. 137–154, 1997.

[3] J. Ashburner and K. Friston, "Unified segmentation," *NeuroImage*, vol. 26, pp. 839–851, 2005.

[4] J. Ashburner and K. J. Friston, "Voxel-based morphometry—The methods," *NeuroImage*, vol. 11, pp. 805–821, 2000.

[5] J. Ashburner, C. Hutton, R. Frackowiak, I. Johnsrude, C. Price, and K. Friston, "Identifying global anatomical differences: Deformation-based morphometry," *Human Brain Mapp.*, vol. 6, pp. 348–357, 1998.

[6] J. Ashburner, P. Neelin, D. L. Collins, A. Evans, and K. Friston, "Incorporating prior knowledge into image registration," *NeuroImage*, vol. 6, no. 4, pp. 344–352, 1997.

[7] A. Ben-Hur, A. Elisseeff, and I. Guyon, "A stability based method for discovering structure in clustered data," in *Pacific Symp. Biocomput.*, 2002, vol. 7, pp. 6–17.

[8] K. K. Bhatia, J. V. Hajnal, B. K. Puri, A. D. Edwards, and D. Rueckert, "Consistent groupwise non-rigid registration for atlas construction," in *IEEE Int. Symp. Biomed. Imag.: Nona to Macro*, 2004, vol. 1, pp. 908–911.

[9] D. Blezek and J. Miller, "Atlas stratification," *Med. Image Anal.*, vol. 11, no. 5, pp. 443–457, 2007.

[10] R. P. Brent, *Algorithms for Minimization Without Derivatives*. Englewood Cliffs, NJ: Prentice-Hall, 1973.

[11] J. R. Clifford, R. C. Petersen, P. C. O'Brien, and E. G. Tangalos, "MR-based hippocampal volumetry in the diagnosis of Alzheimer's disease," *Neurology*, vol. 42, 1992.

[12] D. L. Collins, P. Neelin, T. M. Peters, and A. C. Evans, "Automatic 3-D intersubject registration of MR volumetric data in standardized Talairach space," *J. Comput. Assist. Tomogr.*, vol. 18, no. 2, pp. 192–205, 1994.

[13] D. L. Collins, A. P. Zijdenbos, W. F. C. Baare, and A. C. Evans, "ANI-MAL + INSECT: Improved cortical structure segmentation," in *Proc. Inf. Process. Med. Imag.*, 1999, vol. 1613, pp. 210–223.

[14] M. De Craene, A. B. d Aische, B. Macq, and S. K. Warfield, "Multi-subject registration for unbiased statistical atlas construction," in *Proceedings MICCAI 2004: Medical Image Computing and Computer-Assisted Intervention*. New York: Springer, vol. 3216, Lecture Notes Computer Sciene, pp. 655–662.

[15] C. Davatzikos, A. Gene, D. Xua, and S. M. Resnick, "Voxel-based morphometry using the ravens maps: Methods and validation using simulated longitudinal atrophy," *NeuroImage*, vol. 14, pp. 1361–1369, 2001.

[16] C. DeCarli, J. V. Haxby, J. A. Gillette, D. Teichberg, S. I. Rapoport, and M. B. Schapiro, "Longitudinal changes in lateral ventricular volume in patients with dementia of the alzheimer type," *Neurology*, vol. 42, no. 10, pp. 2029–2036, 1992.

[17] F. E. Harrell, *Regression Modelling Strategies*. New York: Springer, 2001.

[18] B. Fischl, D. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. van der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, A. Montillo, N. Makris, B. Rosen, and A. Dale, "Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain," *Neuron*, vol. 33, no. 3, pp. 341–355, 2002.

[19] B. Fischl, A. van der Kouwe, C. Destrieux, E. Halgren, F. Ségonne, D. Salat, E. Busa, L. J. Seidman, J. Goldstein, D. Kennedy, V. Cavinnes, N. Makris, B. Rosen, and A. Dale, "Automatically parcellating the human cerebral cortex," *Cerebral Cortex*, vol. 14, pp. 11–22, 2004.

[20] A. F. Fotenos, A. Z. Snyder, L. E. Girton, J. C. Morris, and R. L. Buckner, "Normative estimates of cross-sectional and longitudinal brain volume decline in aging and ad," *Neurology*, vol. 64, pp. 1032–1039, 2005.

[21] S. Gao, H. C. Hendrie, K. S. Hall, and S. Hui, "The relationships between age, sex, and the incidence of dementia and alzheimer disease: A meta-analysis," *Arch. General Psychiatry*, vol. 55, pp. 809–815, 1998.

[22] D. S. Geldmacher and P. J. Whitehouse, "Differential diagnosis of alzheimer's disease," *Neurology*, vol. 48, no. 5, 1997.

[23] C. D. Gooda, I. S. Johnsrude, J. Ashburner, R. N. A. Henson, K. J. Friston, and R. S. J. Frackowiak, "A voxel-based morphometric study of ageing in 465 normal adult human brains," *NeuroImage*, vol. 14, no. 1, pp. 21–36, 2001.

[24] M. D. Greicius, G. Srivastava, A. L. Reiss, and V. Menon, "Default-mode network activity distinguishes alzheimer's disease from healthy aging: Evidence from functional MRI," *Proc. Nat. Acad. Sci.*, vol. 101, no. 13, pp. 4637–4642, 2004.

[25] A. Guimond, F. J. Meunier, and J. R. Thirion, "Average brain models: A convergence study," *Comput. Vis. Image Understand.*, vol. 77, no. 2, pp. 192–210, 2000.

[26] D. Head, R. L. Buckner, J. S. Shimony, L. E. Williams, E. Akbudak, T. E. Conturo, M. McAvoy, J. C. Morris, and A. Z. Snyder, "Differential vulnerability of anterior white matter in nondemented aging with minimal acceleration in dementia of the alzheimer type: Evidence from diffusion tensor imaging," *Cerebral Cortex*, vol. 14, pp. 410–423, 2004.

[27] R. A. Heckemann, J. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers, "Automatic anatomical brain MRI segmentation combining label propagation and decision fusion," *NeuroImage*, vol. 33, no. 1, pp. 115–126, 2006.

[28] Y. Hirayasu, M. E. Shenton, D. F. Salisbury, C. C. Dickey, I. A. Fischer, P. Mazzoni, T. Kisler, H. Arakaki, J. S. Kwon, J. E. Anderson, D. Yurgelun-Todd, M. Tohen, and R. W. McCarley, "Lower left temporal lobe MRI volumes in patients with first-episode schizophrenia compared with psychotic patients with first-episode affective disorder and normal subjects," *Am. J. Psychiatry*, vol. 155, no. 10, pp. 1384–1391, 1998.

[29] R. Honea, T. J. Crow, D. Passingham, and C. E. Mackay, "Regional deficits in brain volume in schizophrenia: A meta-analysis of voxel-based morphometry studies," *Am. J. Psychiatry*, vol. 162, pp. 2233–2245, 2005.

[30] L. Ibanez, W. Schroeder, L. Ng, and J. Gates, The ITK Software Guide.. 2005.

[31] S. Joshi, B. Davis, M. Jomier, and G. Gerig, "Unbiased diffeomorphism atlas construction for computational anatomy," *NeuroImage*, vol. 23, pp. 151–160, 2004.

[32] R. Liu, L. Lemieux, G. S. Bell, S. M. Sisodiya, S. D. Shorvon, J. W. A. S. Sander, and J. S. Duncan, "A longitudinal study of brain morphometrics using quantitative magnetic resonance imaging and difference image analysis," *NeuroImage*, vol. 20, pp. 22–33, 2003.

[33] N. Makris, A. J. Worth, A. G. Sorensen, G. M. Papadimitriou, O. Wu, T. G. Reese, V. J. Wedeen, T. L. Davis, J. W. Stakes, V. S. Caviness, E. Kaplan, B. R. Rosen, D. N. Pandya, and D. N. Kennedy, "Morphometry of in vivo human white matter association pathways with diffusion-weighted magnetic resonance imaging," *Ann. Neurol.*, vol. 42, no. 6, pp. 951–962, 1997.

[34] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open access series of imaging studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults," *J. Cognitive Neurosci.*, vol. 19, pp. 1498–1507, 2007.

[35] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, no. 2, pp. 159–179, 1985.

[36] J. C. Morris, "The clinical dementia rating (CDR): Current version and scoring rules," *Neurology*, vol. 43, pp. 2412–2414, 1993.

[37] H. Park, P. H. Bland, A. O. Hero, and C. R. Meyer, "Least biased target selection in probabilistic atlas construction," in *Proceedings MICCAI 2005: Medical Image Computing and Computer-Assisted Intervention*. New York: Springer, 2005, vol. 3750, Lecture Notes Computer Science, pp. 419–426.

[38] K. M. Pohl, J. Fisher, W. Grimson, R. Kikinis, and W. M. Wells, "A bayesian model for joint segmentation and registration," *NeuroImage*, vol. 31, pp. 228–239, 2006.

[39] J. C. Pruessner, L. M. Li, W. Series, M. Pruessner, D. L. Collins, N. Ka-bani, S. Lupien, and A. C. Evans, "Volumetry of hippocampus and amygdala with high-resolution MRI and three-dimensional analysis software: Minimizing the discrepancies between laboratories," *Cerebral Cortex*, vol. 10, no. 4, pp. 433–442, 2000.

[40] E. H. Rubin, M. Storandt, J. R. Miller, D. A. Kinscherf, E. A. Grant, J. C. Morris, and L. Berg, "A prospective study of cognitive function and onset of dementia in cognitively healthy elders," *Arch. Neurol.*, vol. 55, pp. 395–401, 1998.

[41] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: Application to breast MR images," *IEEE Trans. Med. Imag.*, vol. 18, no. 8, pp. 712–721, Aug. 1999.

[42] M. R. Sabuncu, S. K. Balci, and P. Golland, "Discovering modes of an image population through mixture modeling," in *Proceedings MICCAI 2008: Medical Image Computing and Computer-Assisted Intervention*. New York: Springer, 2008, vol. 5242, Lecture Notes Computer Science, pp. 381–389.

[43] R. I. Scahill, C. Frost, R. Jenkins, J. L. Whitwell, M. N. Rossor, and N. C. Fox, "A longitudinal study of brain volume changes in normal aging using serial registered magnetic resonance imaging," *Arch. Neurol.*, vol. 60, pp. 989–994, 2003.

[44] J. B. Schulz, M. Skalej, D. Wedekind, A. R. Luft, M. Abele, K. Voigt, J. Dichgans, and T. Klockgether, "Magnetic resonance imaging-based volumetry differentiates idiopathic parkinson's syndrome from multiple system atrophy and progressive supranuclear palsy," *Ann. Neurol.*, vol. 45, no. 1, pp. 65–74, 2002.

[45] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, 1978.

[46] C. Studholme and V. Cardenas, "A template free approach to volumetric spatial normalization of brain anatomy," *Pattern Recognit. Lett.*, vol. 25, pp. 1191–1202, 2004.

[47] J. Talairach and P. Tornoux, *Co-Planar Stereotaxic Atlas of the Human Brain*. New York: Thieme Medical Publishers, 1998.

[48] P. M. Thompson, R. P. Woods, M. S. Mega, and A. W. Toga, "Mathematical/computational challenges in creating deformable and probabilistic atlases of the human brain," *Human Brain Mapp.*, vol. 9, no. 2, pp. 81–92, 2000.

[49] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *J. R. Stat. Soc.*, ser. B (Statistical Methodology), vol. 63, no. 2, pp. 411–423, 2002.

[50] A. Tsai, W. M. Wells, S. K. Warfield, and A. S. Willsky, "An EM algorithm for shape classification based on level sets," *Med. Image Anal.*, vol. 9, pp. 491–502, 2005.

[51] C. J. Twining, T. Cootes, S. Marsland, V. Petrovic, R. Schestowitz, and C. Taylor, "A unified information-theoretic approach to groupwise non-rigid registration and model building," in *Proceedings Information Processing in Medical Imaging 2005*. New York: Springer, vol. 3565, Lecture Notes Computer Science, pp. 1–14.

[52] L. Wang, J. S. Swank, I. E. Glick, M. H. Gado, M. I. Miller, J. C. Morris, and J. G. Csernansky, "Changes in hippocampal volume and shape across time distinguish dementia of the alzheimer type from healthy aging," *NeuroImage*, vol. 20, no. 2, pp. 667–682, 2003.

[53] R. P. Woods, M. Dapretto, N. L. Sicotte, A. W. Toga, and J. C. Mazziotta, "Creation and use of a Talairach-compatible atlas for accurate, automated, nonlinear intersubject registration, and analysis of functional imaging data," *Human Brain Mapp.*, vol. 8, no. 2–3, pp. 73–79, 1999.

[54] R. P. Woods, S. T. Grafton, J. D. Watson, N. L. Sicotte, and J. C. Mazziotta, "Automated image registration: II Intersubject Validation of Linear and Nonlinear Models," *Camp. Assist. Tomography*, vol. 22, no. 1, pp. 153–165, 1998.

[55] B. T. T. Yeo, M. R. Sabuncu, R. Desikan, B. Fischl, and P. Golland, "Effects of registration regularization and atlas sharpness on segmentation accuracy," *Med. Image Anal.*, vol. 12, no. 5, pp. 603–615, 2008.

[56] B. T. T. Yeo, M. R. Sabuncu, H. Mohlberg, K. Amunts, K. Zilles, P. Golland, and B. Fischl, "What data to co-register for computing atlases," in *Proc. Int. Conf. Comput. Vis., IEEE Comput. Soc. Workshop Math. Methods Biomed. Image Anal.*, 2007, pp. 1–8.

[57] L. Zollei, M. Jenkinson, S. Timonerm, and W. M. Wells, "A marginalized MAP approach and EM optimization for pair-wise registration," in *Proceedings Information Processing in Medical Imaging*. New York: Springer, 2007, vol. 4584, Lecture Notes Computer Science, pp. 662–674.

[58] L. Zöllei, E. Learned-Miller, E. Grimson, and W. Wells, "Efficient population registration of 3-D data," in *Computer Vision for Biomedical Image Applications*. New York: Springer, 2005, vol. 3765, Lecture Notes Computer Science, pp. 291–301.