Encoding Probabilistic Brain Atlases Using Bayesian Inference

Koen Van Leemput, *Member, IEEE*

***Abstract*—This paper addresses the problem of creating probabilistic brain atlases from manually labeled training data. Probabilistic atlases are typically constructed by counting the relative frequency of occurrence of labels in corresponding locations across the training images. However, such an "averaging" approach generalizes poorly to unseen cases when the number of training images is limited, and provides no principled way of aligning the training datasets using deformable registration. In this paper, we generalize the generative image model implicitly underlying standard "average" atlases, using mesh-based representations endowed with an explicit deformation model. Bayesian inference is used to infer the optimal model parameters from the training data, leading to a simultaneous group-wise registration and atlas estimation scheme that encompasses standard averaging as a special case. We also use Bayesian inference to compare alternative atlas models in light of the training data, and show how this leads to a data compression problem that is intuitive to interpret and computationally feasible. Using this technique, we automatically determine the optimal amount of spatial blurring, the best deformation field flexibility, and the most compact mesh representation. We demonstrate, using 2-D training datasets, that the resulting models are better at capturing the structure in the training data than conventional probabilistic atlases. We also present experiments of the proposed atlas construction technique in 3-D, and show the resulting atlases' potential in fully-automated, pulse sequence-adaptive segmentation of 36 neuroanatomical structures in brain MRI scans.**

***Index Terms*—Atlas formation, Bayesian inference, brain modeling, computational anatomy, image registration, mesh generation, model comparison.**

I. Introduction

THE study of many neurodegenerative and psychiatric diseases benefits from fully-automated techniques that are able to reliably assign a neuroanatomical label to each voxel in magnetic resonance (MR) images of the brain. In order to cope with the complex anatomy of the human brain, the large overlap in intensity characteristics between structures of interest, and the dependency of MRI intensities on the acquisition sequence used, state-of-the-art MRI brain labeling techniques rely on prior information extracted from a collection of manually labeled training datasets [1]–[9]. Most typically, this prior information is represented in the form of *probabilistic atlases*, constructed by first registering the training datasets together using affine transformations, and then calculating the probability of each voxel being occupied by a particular structure as the relative frequency that structure occurred at that voxel across the training datasets.

While such "average" atlases are intuitive and straightforward to compute, they are not necessarily the best way to extract population-wise statistics from the training data. A first problem is that probabilistic atlases, built from a limited number of training datasets, tend to generalize poorly to subjects not included in the training database. This is essentially an *overfitting* problem: due to the enormous variability in cortical patterns across individuals, the atlas may erroneously assign a zero probability for observing a particular label at a specific location, simply because that label did not occur at that location in the training datasets. In order to alleviate this problem, a common strategy is to blur probabilistic atlases using e.g., a Gaussian kernel, mimicking the effect of a larger training database (see, for instance, [10] and [5]). While it is intuitively clear that less blurring will be needed as the size of the training database grows, no clear guidelines exist to determine what the optimal amount of blurring is for a given dataset, or when blurring is no longer necessary.

Another problem with "average" atlases is that they do not model nonlinear deformations that would allow one to align corresponding structures across the training datasets, although this would seem a natural way to capture anatomical variations. Furthermore, even if nonlinear deformations were explicitly allowed during the atlas construction phase (as in [11] and [12]), it is not clear how flexible a deformation field model would be appropriate for the task at hand. While the sharpness and structural resolution of population averages after nonrigid alignment is a typical measure of success in intersubject registration of neuroanatomical images [13]–[19], such results are not necessarily helpful in building priors: more flexible deformation fields will always allow us to align the training datasets better, but are also much weaker at representing the *typical* variations observed across the population.

In this paper, we propose several advancements to the probabilistic atlas construction problem, providing quantitative answers to the issues raised above. Central to our approach is the notion that standard probabilistic atlases implicitly assume a specific *generative image model* for the training datasets at hand, and that estimating the relative frequency of occurrence of various structures in each voxel is, in fact, a Bayesian assessment of the most likely parameters of this model given the training data. With this Bayesian modeling framework in mind, the novel contribution in this paper is three-fold.

Manuscript received August 06, 2008; revised October 31, 2008. First published December 09, 2008; current version published May 28, 2009. This work was supported in part by the National Center for Research Resources (P41-RR14075, the BIRN Morphometric Project BIRN002, U24 RR021382, and the Neuroimaging Analysis Center. NAC P41-RR13218), in part by the National Institute for Biomedical Imaging and Bioengineering (R01EB006758 and the National Alliance for Medical Image Analysis, NAMIC U54-EB005149), in part by the National Institute for Neurological Disorders and Stroke (R01 NS052585-01 and R01-NS051826), and in part by the NSF CAREER 0642971 Award and the Autism & Dyslexia Project funded by the Ellison Medical Foundation.

The author is with the Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Charlestown, MA 02129 USA and also with the MIT Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 01239 USA (e-mail: koen@nmr.mgh.harvard.edu).

Digital Object Identifier 10.1109/TMI.2008.2010434

1) We propose a generalization of the generative image model underlying traditional probabilistic atlases, using a mesh-based atlas representation, and allowing for nonlinear deformations. Using the notation H for a specific model and θ for the parameters of such a model, alternative models H_i are fully described by a prior distribution $p(\theta \,|\, H_i)$ for their model parameters; and a likelihood distribution $p(D \,|\, \theta, H_i)$ that defines what predictions the model makes about the training data D. In the context of this paper, different models H_i refer to different mesh configurations and/or different values for a hyper-parameter regulating the flexibility of the deformation field models; the parameters θ parametrize the deformation fields and the relative frequency of occurrence of structures at various locations throughout the atlas.
2) Assuming that a given model H_i is true, we use Bayes' theorem to try to infer what the model's parameters θ may be, given the data D. Maximizing

$$p(\theta \,|\, D, H_i) \propto p(D \,|\, \theta, H_i) p(\theta \,|\, H_i)$$

leads to a novel *group-wise registration* process [16]–[22], in which the deformations warping the atlas to each of the training datasets are estimated simultaneously with an unbiased probabilistic atlas. For a specific choice of model H_i, this process devolves into the standard "average" probabilistic atlas estimation.
3) Again using Bayes' rule, we compare various alternative models H_i in light of the training data D, by evaluating

$$p(H_i \,|\, D) \propto p(D \,|\, H_i) p(H_i).$$

Having no *a priori* preference for any model H_i over the others, we use equal priors $p(H_i)$ for alternative models, and use the so-called *evidence* $p(D \,|\, H_i)$ to rank them. This allows us to objectively assess the optimal amount of blurring in a probabilistic atlas for given training data, to determine the optimal flexibility of deformation field models, and to construct compact atlas representations using content-adaptive meshes.

To the best of our knowledge, the *atlas model comparison* problem (item 3) has not been addressed before in the literature, so let us briefly point out the intuition behind our Bayesian approach (see [23] for an excellent introduction to Bayesian model comparison). The key observation is that ranking alternative models according to their evidence automatically and quantitatively safeguards us from using over-parametrized models that would constitute poor priors. As an example, consider a model that allows exceedingly flexible deformations of the atlas. While such a model can be fitted extremely well to the training data, its evidence, defined as

$$p(D \,|\, H_i) = \int_\theta p(D \,|\, \theta, H_i) p(\theta \,|\, H_i) \mathrm{d}\theta$$

is very low: because the range of possible outcomes is so large, the probability of observing exactly the training data D must be very low. Indeed, it would be quite a *coincidence* that, if we drew samples from such an underconstrained model, the results would happen to look like brains!

Another way to gain insight into how Bayesian model comparison works, is to write the evidence down in terms of the length, measured in bits, of the shortest message that communicates the training data without loss to a receiver when a certain model H_i is used. Following Shannon theory, this length is $-\log_2 p(D \,|\, H_i)$; searching for a model that maximizes the evidence is thus equivalent to trying to discover regularities in the training data, allowing us to maximally *compress* it. Note that nothing is said about encoding at the optimal parameters; intuitively, these parameter values will need to be encoded somehow as well, automatically safeguarding against overly complex models with too many free parameters.

In this paper, we only address the problem of learning, from manually labeled training data, a prior distribution that makes predictions about where neuroanatomical labels typically occur throughout images of new subjects. Once built, such a prior can be freely mixed and matched with a variety of probabilistic atlas-based modeling and optimization techniques to obtain automated segmentations of brain MRI data [1], [2], [4], [5], [9], [24], [25]. We note that this concept of probabilistic atlases is different from the one in which structure-specific intensity distributions are learned simultaneously with the prior as well [7], [26].

This paper is structured as follows. Section II introduces our generalized atlas model. In Section III, we describe three levels of Bayesian inference, derive practical optimizers and approximations, and interpret the inference problem in terms of message encoding using binary strings. Sections IV and V report, respectively, experiments and results on manually labeled datasets in 2-D. In Section VI, we present experiments of the proposed atlas construction technique in 3-D, and show the resulting atlases' potential in fully-automated, pulse sequence-adaptive segmentation of 36 neuroanatomical structures. Finally, we relate our approach to existing work and present a future outlook in Section VII. An early version of this work was presented in [27].

II. Generative Image Model

The techniques proposed in this paper apply equally well in the 2-D domain, using triangular atlas mesh representations, as in the 3-D domain, using tetrahedral meshes. For ease of presentation, we will use triangular meshes throughout the theoretical sections, keeping in mind that the described procedures have their direct equivalent in tetrahedral meshes as well.

Let there be M manually labeled images $L_m, m = 1, 2, \ldots, M$. Each image $L_m = \{l_i^m, i = 1, 2, \ldots, I\}$ has a total of I pixels, with $l_i^m \in \{1, 2, \ldots, K\}$ denoting the one of K possible labels assigned to pixel i. We model these images (and subsequent ones that are to be analyzed) as being generated by the following process:

1) First, a triangular mesh covering the whole image domain is constructed, defined by the position of its N mesh nodes $\boldsymbol{x}^r = \{\boldsymbol{x}_n^r, n = 1, 2, \ldots, N\}$ and by a *simplicial complex* (a collection of points, line segments, and triangles [28]) $\mathcal{K}$ specifying the mesh connectivity. For the remainder of the paper, we will refer to $\boldsymbol{x}^r$ as the *reference position* of the mesh.
2) A set of label probabilities $\boldsymbol{\alpha}_n = \{\alpha_n^1, \alpha_n^2, \ldots, \alpha_n^K\}$, satisfying $\alpha_n^k \geq 0$ and $\sum_k^K \alpha_n^k = 1$, is assigned to each mesh node, defining how frequently each label tends to occur around that node. In typical probabilistic brain atlases, no more than three labels have a nonzero probability simultaneously at any given location (although these labels vary between locations). Assuming that label proba-

bilities are assigned to each mesh node independently, and letting $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_N\}$ denote the total set of label probabilities of all mesh nodes, we therefore use the prior $p(\boldsymbol{\alpha}) = \prod_n p(\boldsymbol{\alpha}_n)$ with

$$p(\boldsymbol{\alpha}_n) \propto \begin{cases} 0, & \text{if more than 3 labels have} \\ & \text{a nonzero probability} \\ 1, & \text{otherwise.} \end{cases}$$

3) M deformed atlas meshes are obtained by sampling M times from a Markov random field (MRF) prior regulating the position of the mesh nodes[1]:

$$p(\boldsymbol{x} \,|\, \beta, \boldsymbol{x}^r, \mathcal{K}) \propto \exp\left(-\frac{U(\boldsymbol{x} \,|\, \boldsymbol{x}^r, \mathcal{K})}{\beta}\right)$$

with

$$U(\boldsymbol{x} \,|\, \boldsymbol{x}^r, \mathcal{K}) = \sum_{t=1}^{T} U_t^{\mathcal{K}}(\boldsymbol{x} \,|\, \boldsymbol{x}^r). \tag{1}$$

In (1), $U_t^{\mathcal{K}}(\boldsymbol{x} \,|\, \boldsymbol{x}^r)$ is a penalty for deforming triangle t from its shape in the reference position $\boldsymbol{x}^r$, $U(\boldsymbol{x} \,|\, \boldsymbol{x}^r, \mathcal{K})$ is an overall deformation penalty obtained by summing the contributions of all T triangles in the mesh, and the parameter β controls the flexibility of the resulting deformation field prior. In order to insure that the prior is topology preserving, the penalty needs to go to infinity if the Jacobian determinant of any triangle's deformation approaches zero. In this paper, we have used the penalty proposed by Ashburner *et al.* in [21], which has this property; details are given in Appendix A. Note, however, that other definitions would also be possible (such as for instance [30]).

4) From each deformed atlas mesh with position $\boldsymbol{x}^m$, a label image L_m is generated by interpolating the label probabilities at the mesh nodes over the whole image domain, and sampling from the resulting probabilities. Given a mesh with position $\boldsymbol{x}$, the probability of having label k in a pixel i with location $\boldsymbol{x}_i$ is modeled by

$$p_i(k \,|\, \boldsymbol{\alpha}, \boldsymbol{x}, \mathcal{K}) = \sum_{n=1}^{N} \alpha_n^k \phi_n(\boldsymbol{x}_i). \tag{2}$$

In (2), $\phi_n(\cdot)$ denotes an interpolation basis function attached to mesh node n that has a unity value at the position of the mesh node, a zero value at the outward edges of the triangles connected to the node and beyond, and a linear variation across the face of each triangle (see Fig. 1). As a result, the probability of observing a certain label k is given by the label probabilities α_n^k at the mesh nodes, and varies linearly in between the nodes. To complete our model, we assume conditional independence of the labels between pixels given the mesh parameters, so that we have

$$p(L \,|\, \boldsymbol{\alpha}, \boldsymbol{x}, \mathcal{K}) = \prod_{i=1}^{I} p_i(l_i \,|\, \boldsymbol{\alpha}, \boldsymbol{x}, \mathcal{K}) \tag{3}$$

for the probability of seeing label image L.

[1]For simplicity, we will ignore boundary conditions throughout the theoretical sections of the paper. Sliding boundary conditions [29] were used, in which mesh nodes lying on an image edge can only slide along that edge.

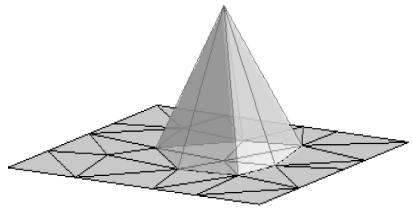

Fig. 1. In the generative model, label probabilities are interpolated from the probabilities in the mesh nodes using a linear combination of interpolation basis functions $\phi_n(\cdot)$. This figure shows the interpolation basis function for one mesh node: it varies linearly over the face of each triangle attached to the node, and has only limited, local support.

III. Bayesian Inference

A. First Level of Inference

Given manually labeled training data in the form of M label images $L_m, m = 1, 2, \ldots, M$, we can infer what the label probabilities and the positions of the mesh nodes in each of the labelings may be. In a Bayesian setting, assessing the Maximum A Posteriori (MAP) parameters $\{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{x}}^1, \ldots, \hat{\boldsymbol{x}}^M\}$ involves maximizing

$$\prod_{m=1}^{M} [p(L_m \,|\, \boldsymbol{\alpha}, \boldsymbol{x}^m, \mathcal{K}) p(\boldsymbol{x}^m \,|\, \beta, \boldsymbol{x}^r, \mathcal{K})] p(\boldsymbol{\alpha}), \tag{4}$$

which is equivalent to minimizing

$$\sum_{m=1}^{M} [-\log p(L_m \,|\, \boldsymbol{\alpha}, \boldsymbol{x}^m, \mathcal{K}) - \log p(\boldsymbol{x}^m \,|\, \beta, \boldsymbol{x}^r, \mathcal{K})] - \log p(\boldsymbol{\alpha}). \tag{5}$$

We alternately optimize the label probabilities in the mesh nodes $\boldsymbol{\alpha}$, keeping the position parameters fixed, and update each of the positions $\boldsymbol{x}^m$ while keeping the label probabilities fixed. Optimizing the positions is a registration process, bringing each of the training samples in spatial correspondence with the current atlas. The gradient of (5) with respect to $\boldsymbol{x}^m$ is given in analytical form through (1) and (2), and we perform this registration by global gradient descent (although we also use a local node-by-node optimization in specific circumstances; see later).

Assessing the optimal label probabilities in the mesh nodes for a given registration of the training samples can be done iteratively using an expectation-maximization (EM) algorithm [31]. We initialize the algorithm with label probabilities in which all labels are equally alike in all mesh nodes. At each iteration, we then construct a lower bound to (4) that touches (4) at the current values of $\boldsymbol{\alpha}$

$$\prod_{m=1}^{M} \left[\prod_{i=1}^{I} \prod_{n=1}^{N} \left(\frac{\alpha_n^{l_i^m} \phi_n^m(\boldsymbol{x}_i)}{W_{i,n}^m} \right)^{W_{i,n}^m} p(\boldsymbol{x}^m \,|\, \beta, \boldsymbol{x}^r, \mathcal{K}) \right] p(\boldsymbol{\alpha}). \tag{6}$$

In (6), the weights

$$W_{i,n}^m = \frac{\alpha_n^{l_i^m} \phi_n^m(\boldsymbol{x}_i)}{\sum_{n'=1}^{N} \alpha_{n'}^{l_i^m} \phi_{n'}^m(\boldsymbol{x}_i)}$$

associate each pixel in each example with each of the mesh nodes; note that, due to the limited support of the basis functions $\phi_n^m(\cdot)$, a pixel's weights can only be nonzero for the

three mesh-nodes attached to the triangle containing it. Once the lower bound is constructed, optimizing it with respect to the label probabilities is straightforward: each node's label probabilities are obtained as the relative frequency of occurrence of the labels in the pixels assigned to it[2]:

$$\alpha_n^k \leftarrow \frac{\sum_{m=1}^{M}\sum_{i=1}^{I} W_{i,n}^m \, \delta_{l_i^m,k}}{\sum_{m=1}^{M}\sum_{i=1}^{I} W_{i,n}^m} \quad \forall n,k.$$

With these updated label probabilities, a new lower bound is constructed by recalculating the assignments $W_{i,n}^m$ etc., until convergence. Note that the constraint of maximum three labels with nonzero probability in each node, as dictated by the prior $p(\boldsymbol{\alpha})$, is not explicitly enforced in this algorithm. However, it is easily verified that this condition is automatically fulfilled in practice.

Note that traditional "average" atlases are a special case of the aforementioned EM algorithm: in a regular triangular mesh with no deformations allowed (i.e., $\beta = 0$), where there is a node coinciding exactly with each pixel, the algorithm devolves into a noniterative process that exclusively assigns each pixel to its corresponding mesh node only, resulting in a pixel-wise average of the label images as the MAP estimates for $\boldsymbol{\alpha}$.

B. Second Level of Inference

The results of the atlas parameter estimation scheme described in Section III-A depend heavily on the choice of the hyper-parameter β regulating the flexibility of the deformation fields. Having no prior knowledge regarding the "correct" value of β, we may assign it a flat prior. Using the Bayesian framework, we can then assess its MAP value $\hat{\beta}$ by maximizing

$$p(L_1,\ldots,L_M \,|\, \beta, \boldsymbol{x}^r, \mathcal{K}) = \int_{\boldsymbol{\alpha}} \left(\prod_{m=1}^{M} p(L_m \,|\, \boldsymbol{\alpha}, \beta, \boldsymbol{x}^r, \mathcal{K}) \right) p(\boldsymbol{\alpha}) \mathrm{d}\boldsymbol{\alpha} \quad (7)$$

where

$$p(L_m \,|\, \boldsymbol{\alpha}, \beta, \boldsymbol{x}^r, \mathcal{K}) = \int_{\boldsymbol{x}^m} p(L_m \,|\, \boldsymbol{\alpha}, \boldsymbol{x}^m, \mathcal{K}) p(\boldsymbol{x}^m \,|\, \beta, \boldsymbol{x}^r, \mathcal{K}) \mathrm{d}\boldsymbol{x}^m.$$

Assuming that $p(L_m \,|\, \boldsymbol{\alpha}, \boldsymbol{x}^m, \mathcal{K}) p(\boldsymbol{x}^m \,|\, \beta, \boldsymbol{x}^r, \mathcal{K})$ has a peak at a position $\boldsymbol{x}_\alpha^m$, we may approximate $p(L_m \,|\, \boldsymbol{\alpha}, \beta, \boldsymbol{x}^r, \mathcal{K})$ using Laplace's method, i.e., by locally approximating the integrand by an unnormalized Gaussian. Ignoring interdependencies between neighboring mesh nodes in the Gaussian's covariance matrix, and approximating the prior $p(\boldsymbol{x}^m \,|\, \beta, \boldsymbol{x}^r, \mathcal{K})$ using the pseudo-likelihood approximation [32] and a local Laplace approximation in each node, we obtain[3] (see Appendix B; an illustration is shown in Fig. 2)

$$p(L_m \,|\, \boldsymbol{\alpha}, \beta, \boldsymbol{x}^r, \mathcal{K}) \simeq p(L_m \,|\, \boldsymbol{\alpha}, \boldsymbol{x}_\alpha^m, \mathcal{K}) \cdot \prod_{n=1}^{N} O_n^m \quad (8)$$

[2]Here, $\delta_{k,l}$ denotes the Kronecker delta.

[3]Here, D_θ^2 denotes a matrix of second derivatives, or Hessian.

with

$$O_n^m = \exp\left(-\frac{U(\boldsymbol{x}_\alpha^m \,|\, \boldsymbol{x}^r, \mathcal{K}) - U(\boldsymbol{x}_\alpha^{m\,|\,n} \,|\, \boldsymbol{x}^r, \mathcal{K})}{\beta} \right) \times \sqrt{\frac{\det(\boldsymbol{J}_n^m)}{\det(\boldsymbol{I}_n^m)}}$$

where

$$\boldsymbol{I}_n^m = D_{\boldsymbol{x}_n}^2 [-\log p(L_m \,|\, \boldsymbol{\alpha}, \boldsymbol{x}, \mathcal{K}) - \log p(\boldsymbol{x} \,|\, \beta, \boldsymbol{x}^r, \mathcal{K})]|_{\boldsymbol{x}=\boldsymbol{x}_\alpha^m}$$

and

$$\boldsymbol{J}_n^m = D_{\boldsymbol{x}_n}^2 [-\log p(\boldsymbol{x} \,|\, \beta, \boldsymbol{x}^r, \mathcal{K})]|_{\boldsymbol{x}=\boldsymbol{x}_\alpha^{m\,|\,n}}.$$

Here, $\boldsymbol{x}_\alpha^{m\,|\,n}$ denotes the set of mesh positions that is identical to $\boldsymbol{x}_\alpha^m$ except for the position of node n, which is replaced by the position that maximizes the prior $p(\boldsymbol{x} \,|\, \beta, \boldsymbol{x}^r, \mathcal{K})$ when the positions of all other mesh nodes are fixed to their value in $\boldsymbol{x}_\alpha^m$. Note that calculating this optimal node position, as well as evaluating the factors O_n^m, only involves those triangles that are directly attached to the node under investigation; we use a Levenberg–Marquardt algorithm to carry out the actual optimization.

Plugging (8) into (7), and approximating the factors O_n^m by their values at $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}$, denoted by $\hat{O}_n^m$, we obtain

$$p(L_1,\ldots,L_M \,|\, \beta, \boldsymbol{x}^r, \mathcal{K}) \simeq \prod_{m=1}^{M}\prod_{n=1}^{N} \hat{O}_n^m \cdot \int_{\boldsymbol{\alpha}} \left(\prod_{m=1}^{M} p(L_m \,|\, \boldsymbol{\alpha}, \hat{\boldsymbol{x}}^m, \mathcal{K}) \right) p(\boldsymbol{\alpha}) \mathrm{d}\boldsymbol{\alpha}.$$

The remaining integral cannot, in general, be obtained analytically. To sidestep this difficulty, we replace $p(L_m \,|\, \boldsymbol{\alpha}, \hat{\boldsymbol{x}}^m, \mathcal{K})$ by the lower bound

$$\prod_{i=1}^{I}\prod_{n=1}^{N} \left(\frac{\alpha_n^{l_i^m} \phi_n^m(\boldsymbol{x}_i)}{\hat{W}_{i,n}^m} \right)^{\hat{W}_{i,n}^m}$$

used in the EM algorithm of Section III-A, which touches $p(L_m \,|\, \boldsymbol{\alpha}, \hat{\boldsymbol{x}}^m, \mathcal{K})$ at the optimal label probabilities $\hat{\boldsymbol{\alpha}}$. Taking into account the prior $p(\boldsymbol{\alpha})$, which only allows nonzero probabilities for three labels simultaneously in each node but is otherwise flat, and using Stirling's approximation for the Gamma function $\Gamma(x+1) \simeq x^x e^{-x}$, we finally obtain (see Appendix C)

$$p(L_1,\ldots,L_M \,|\, \beta, \boldsymbol{x}^r, \mathcal{K}) \simeq \prod_{m=1}^{M}\prod_{n=1}^{N} \hat{O}_n^m \cdot \prod_{n=1}^{N} \hat{R}_n \cdot \prod_{m=1}^{M} p(L_m \,|\, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{x}}^m, \mathcal{K}) \quad (9)$$

with

$$\hat{R}_n = \binom{K}{3} \cdot \frac{2!\Gamma(\hat{N}_n+1)}{\Gamma(\hat{N}_n+3)} = \frac{12}{(K-2)(K-1)K(\hat{N}_n+1)(\hat{N}_n+2)}$$

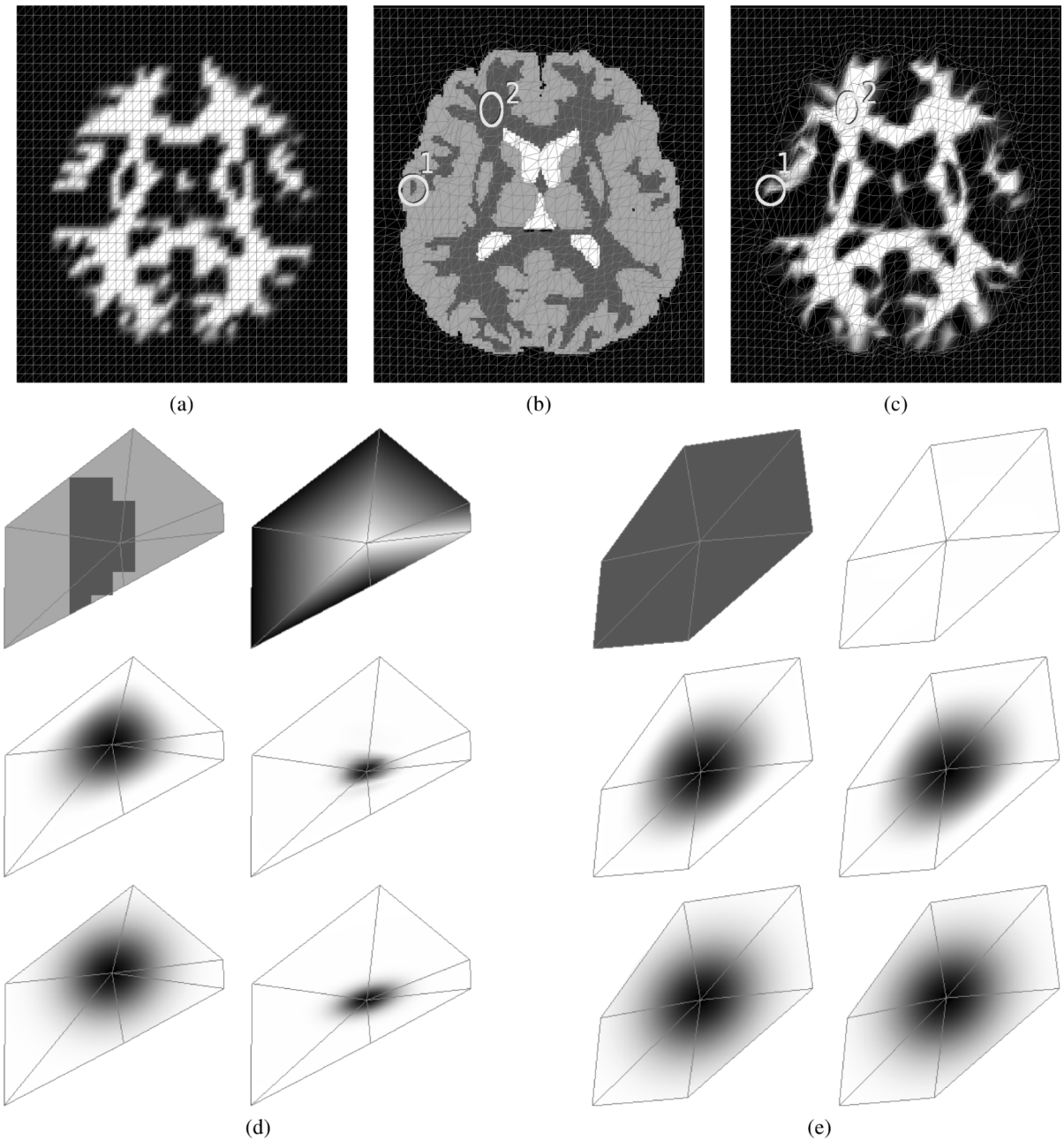

Fig. 2. The probability of seeing a label image L (b) given an atlas in its reference position (a) is obtained by multiplying the probability of seeing the label image given the optimally deformed atlas (c) with a factor with magnitude less than one, which is a penalty for not actually knowing the deformation field. In the illustration of the atlases, white and black indicate a white matter probability of 1 and 0, respectively. We approximate the penalty factor for not knowing the optimal deformation field by a product of local penalties O_n, one for each mesh node n. Images (d) and (e) illustrate how this local penalty factor is calculated for the node indicated with number 1 and 2 in images (b) and (c), respectively. The top rows in (d) and (e) provide a magnified view of the local neighborhood around the node under investigation in the label image and the deformed atlas. The left and right images in the middle rows show respectively the prior (before any data is seen) and the posterior (after the data in the top left arrives) distributions of the location of the mesh node. Here, dark indicates high probability density values. Finally, the bottom rows show Gaussian approximations to the priors and posteriors of the middle rows that are used to actually calculate the penalty factors. Each node's penalty factor essentially quantifies the difference between the prior and the posterior, by comparing each distribution's MRF energy at the optimal mesh node location and the spread of its Gaussian approximation (see text). As a result, the node shown in (d) incurs a much higher penalty ($O_n \ll 1$) than the node of (e) ($O_n \approx 1$) for not knowing its optimal location. Stated from a data compression point of view, encoding the position of the mesh node requires a high number of bits $-\log_2 O_n$ in (d), but ≈ 0 bits in (e). This reflects the fact that, in contrast to the situation in (e), the position of the node in (d) must be encoded with high precision, because small deviations from its optimal value will result in a large increase in the number of bits required to subsequently encode the labels [top left of (d)]. Note that in reality, the label probabilities in each mesh node are not known either, which gives rise to another penalty factor R_n in each node (see text).

where $\hat{N}_n = \sum_{m=1}^{M} \sum_{i=1}^{I} \hat{W}_{i,n}^m$ denotes the total number of pixels associated with node n at the MAP parameters $\{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{x}}^1, \ldots, \hat{\boldsymbol{x}}^M\}$. Equipped with (9), the MAP estimate $\hat{\beta}$ can be assessed using a line search algorithm (see later).

C. Third Level of Inference

We have assumed so far that the connectivity $\mathcal{K}$ and the reference position $\boldsymbol{x}^r$ of the atlas mesh are known beforehand.

Using the Bayesian framework, however, we can assign objective preferences to alternative models. Having no *a priori* reason to prefer one model over the other, we can rank alternatives based on their likelihood $p(L_1,\ldots,L_M \mid \boldsymbol{x}^r,\mathcal{K}) = \int_\beta p(L_1,\ldots,L_M \mid \beta,\boldsymbol{x}^r,\mathcal{K})\; p(\beta)\; \mathrm{d}\beta$, which can be approximated, using Laplace's method, by

$$\left(\sqrt{2\pi}\; p(\hat{\beta}) \Big/ \sqrt{\tfrac{\partial^2}{\partial\beta^2}[-\log p(L_1,\ldots,L_M \mid \beta,\boldsymbol{x}^r,\mathcal{K})]|_{\beta=\hat{\beta}}}\right) \cdot p(L_1,\ldots,L_M \mid \hat{\beta},\boldsymbol{x}^r,\mathcal{K}).$$

Since changes in the first factor are overwhelmed by changes in the second one, we will ignore the first factor and compare alternative models based on (9), evaluated at the MAP estimate $\hat{\beta}$.

D. Description Length Interpretation

Given that we use (9) both to assess the optimal deformation field flexibility and optimal mesh representations, it is instructive to write it down in terms of the bit length of the shortest message that communicates the training data without loss to a receiver when a certain model is used. Taking the binary logarithm, negating, and rearranging terms, we have

$$-\sum_{n=1}^{N}\log_2 \hat{R}_n - \sum_{m=1}^{M}\sum_{n=1}^{N}\log_2 \hat{O}_n^m - \sum_{m=1}^{M}\log_2 p(L_m \mid \hat{\boldsymbol{\alpha}},\hat{\boldsymbol{x}}^m,\mathcal{K}).$$

According to the three terms, such a message can be imagined as being subdivided into three blocks. Prior to starting the communication, the transmitter estimates the MAP estimates $\{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{x}}^1,\ldots,\hat{\boldsymbol{x}}^M\}$ as laid out in Section III-A. It then sends a message block that encodes the label probabilities in each mesh node (first term). Subsequently, a message block is sent that encodes, for each label image, the position of each mesh node (second term). Finally, the actual data can be encoded using the model at the MAP parameter estimates (third term). From this interpretation, it is clear that finding good models involves balancing the number of bits required to encode the parameters of the model with the number of bits required to encode the training data once those model parameters are known. Overly complex models, while providing a short description of the training data, require an overly lengthy description of their parameters and are automatically penalized.

IV. Experiments

A. Training Data

We evaluated the performance of competing atlas models on 2-D training data, derived from manual annotations that are publicly available at the Internet Brain Segmentation Repository (IBSR) [33]. A first dataset consists of corresponding coronal slices in 18 subjects with delineations of white matter, cerebral cortex, lateral ventricle, caudate, putamen, and accumbens area in both hemispheres (see Fig. 4). Axial slices of the same subjects, containing manual labels of global white matter, gray matter, CSF, and background, constitute a second training dataset (available as supplementary material at http://ieeexplore.ieee.org). Both datasets were obtained by coregistering the annotated volumes of all subjects to the first subject using a 3-D affine registration algorithm [34] and resampling using nearest-neighbor interpolation. The image size of all 2-D slices was 161×145.

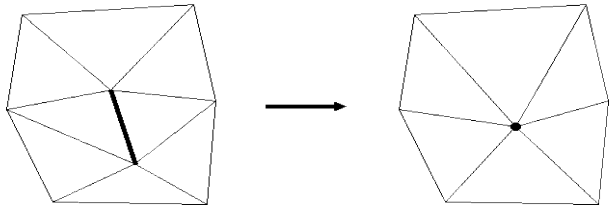

Fig. 3. A mesh can be simplified by unifying two adjacent mesh nodes into a single node using a so-called *edge collapse* operation.

B. Atlas Construction and Comparison

The description length allows us to compare different atlas models in light of the data. On both training datasets, we compared the following models.

- **Full-resolution, nondeformable atlases**. Here, no deformation is allowed ($\beta = 0$), and the atlas mesh is defined on a regular, high-resolution mesh in which each node coincides exactly with the corresponding pixel centers in the training data. This corresponds to the standard notion of probabilistic atlases.
- **Optimal-resolution, nondeformable atlases**. This is similar to standard probabilistic atlases, except that the resolution of the regular mesh is reduced, so that each triangle in the mesh stretches over a number of pixels in the training data.
- **Content-adaptive, nondeformable atlases**. Again, no deformations are allowed ($\beta = 0$), but the mesh nodes are placed strategically so as to obtain a maximally concise representation (see below).
- **Content-adaptive, optimally-deformable atlases**. In addition to seeking the optimal mesh representation, the optimal deformation flexibility β is explicitly assessed as well.

The latter atlas model involves a joint estimation of both the optimal deformation flexibility and the optimal mesh connectivity, which poses a very challenging optimization problem. For our experiments, we have used the following three-step scheme, which is in no way optimal but which yields useful answers in a practically feasible fashion.

First, the model parameters of a high-resolution, regular mesh-based atlas were estimated for a given, fixed value of the deformation flexibility, using the scheme described in Section III-A ($\beta = 10$ was used in our experiments). The parameter estimation proceeded in a four-level multiresolution fashion, in which first a low-resolution mesh was fitted, which was then upsampled and fitted again, etc., until the mesh reached the full resolution of the training data.

Second, a mesh simplification procedure [35], [36] was employed that repeatedly visits each edge in the mesh at random, and compares the effect on the description length of either keeping the edge while optimizing the reference position of the two nodes attached to it, or collapsing the edge and optimizing the reference position of the resulting unified node (see Fig. 3). The optimization of the reference positions of the

Fig. 4. First training dataset: corresponding coronal slices with 13 labels in 18 subjects. Only the data of the first three subjects are shown.

mesh nodes was performed using Powell's direction set [37], involving for each trial reference position an inner optimization of the atlas' label probabilities and deformations as described in Section III-A. Since each edge operation tends to change the resulting atlas only locally, this inner optimization was restricted to a small area around the edge under investigation: only the model parameters in the nodes directly affected were updated in the experiments reported here.

Finally, the optimal β was assessed for the resulting content-adaptive, deformable atlas using a line search. For each trial value for β, the atlas model was refitted according to Section III-A, and the description length was evaluated. Since it is imperative that the atlas model parameters are optimized properly in order to accurately reflect the effect of small changes in β on the resulting description length, the global gradient-descent registration component of Section III-A was replaced by an Iterated conditional modes (ICMs) scheme [38], in which all nodes in each of the training images are repeatedly visited, and individually optimized using a Levenberg–Marquardt algorithm, keeping the position of all other nodes fixed.

The atlas encoding scheme took 14 h for each training dataset on an AMD Opteron 275 processor, with almost all time consumed by the mesh simplification step. For the content-adaptive, nondeformable atlas meshes, the same mesh simplification procedure was used as in the deformable case, but it was much less computationally demanding there as the inner optimization is only over the atlas label probabilities: the mesh node positions in each training dataset can simply be copied from the reference positions.

C. Visualization of the Results

In addition to quantitatively evaluating competing atlas models by comparing their message length, we can also explore what aspects of the data they have actually captured by synthesizing samples from the probability distributions that they describe. Following the generative image model of Section II, generating samples involves sampling from the deformation field model, interpolating the deformed atlases at the pixel locations, and assigning an anatomical label to each pixel accordingly. We sampled from our deformation field model using a Markov Chain Monte Carlo (MCMC) technique known as the Hamiltonian Monte Carlo method [39], which is more efficient than traditional Metropolis schemes because it uses gradient information to reduce random walk behavior [23]. In a nutshell, the method generates samples from our MRF prior by iteratively assigning an artificial, random momentum to each mesh node, and simulating the dynamics of the resulting system for a certain amount of time, where the MRF energy acts as an internal force on the mesh nodes.

V. Results

Results for the first training dataset (Fig. 4) are presented in Figs. 5–7. Results for the second dataset are qualitatively similar, and are available as supplementary material at http://ieeexplore.ieee.org.

Considering the first training dataset, Fig. 5(a) shows the *full-resolution, nondeformable atlas* built from the 18 training images. The figure also contains a schematic representation of the data encoding message. Since no deformations are involved, only the label probabilities in each pixel location and the residual data uncertainty need to be encoded (former and latter message block, respectively). In absolute terms, the message length is ≈ 549 kbits; this should be compared to a literal description of the data in which the one out of 13 possible labels in each of the 161×145 pixels in the 18 training images is described by $\log_2(13)$ bits, yielding a message length of ≈ 1.482 Mb. In other words, the probabilistic atlas representation clearly captures some of the regularities in the training data, allowing it to be compressed by approximately 64% compared to when the pixel labels are assumed to be completely random. Nevertheless, the majority of the bits are spent encoding the model parameters, in this case the label probabilities. This indicates that we may be overfitting the training data, limiting the atlas' ability to predict unseen cases.

If we gradually increase the distance between the mesh nodes in both directions, we obtain increasingly sparse mesh representations in which the number of bits spent encoding the model parameters decreases, at the expense of longer data encoding blocks [see Fig. 6(a)]. The optimal distance between the mesh nodes, yielding the shortest overall message length, is around 5.5 times the pixel distance, resulting in a mesh with approximately 30 times less nodes than the full-resolution atlas. The resulting *optimal-resolution, nondeformable atlas* is depicted in Fig. 5(b); its message length is around 42% of that of the

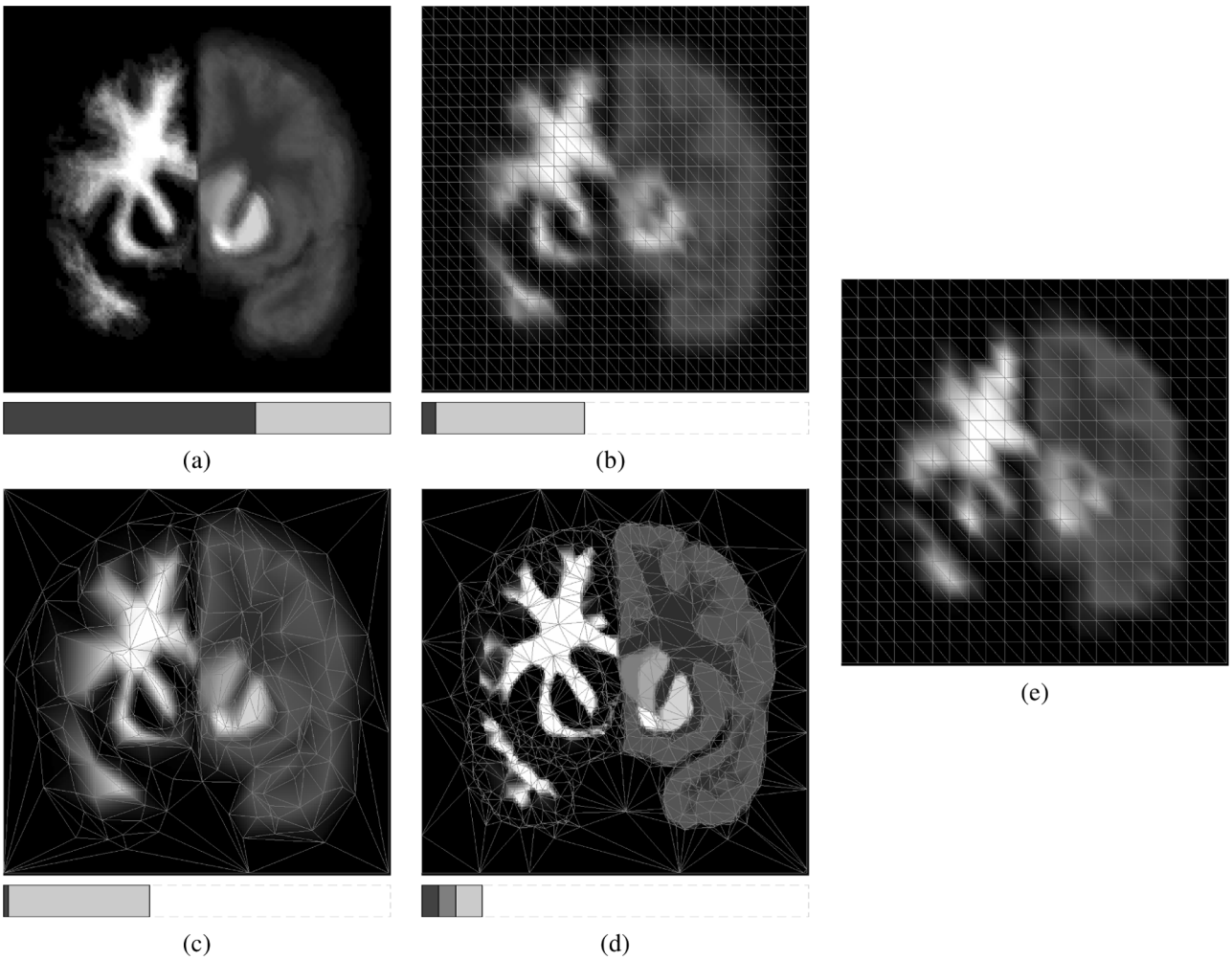

Fig. 5. Competing atlas models constructed from the first training dataset (see Fig. 4): *full-resolution, nondeformable atlas* (a), *optimal-resolution, nondeformable atlas* (b), *content-adaptive, nondeformable atlas* (c), and *content-adaptive, optimally-deformable atlas* (d). Also shown is the *optimal-resolution, nondeformable atlas* constructed using only 3 out of the 18 training images (e), which should be compared to (b). White and black indicate a white matter probability of 1 and 0, respectively. The right side of the brain has been color-coded in the atlases for visualization purposes. Under each atlas (a)–(d) is depicted a schematic view of the shortest message that encodes the training data: dark gray indicates the label probabilities message block, intermediate gray represents the node position message block, and light gray stands for the data message block. All message lengths are represented relative to the message length of the *full-resolution, nondeformable atlas* [image (a)], which in itself already provides a 64 % compression rate (see text for more details).

full-resolution atlas. Note that the lower mesh resolution necessarily introduces a certain amount of blur in the resulting atlas, thereby improving its generalization ability.

At this point, we may wonder how the optimal mesh resolution is affected when the number of training images used to build the atlas is altered. Intuitively, the risk of overfitting is higher when less training data is available, and the optimal mesh resolution should go down accordingly. We can verify that this is indeed the case: Fig. 5(e) shows the optimal-resolution atlas when only 3 training images are used, as opposed to 18. Compared to the atlas of Fig. 5(b), the number of mesh nodes is further decreased by another 47%. Note that using a lower mesh resolution is akin to increasing the amount of blur in the resulting atlas; Bayesian inference thus automatically and quantitatively determines the "correct" amount of blurring that should be applied.

Returning back to 18 training images, Fig. 5(c) shows the *content-adaptive, nondeformable atlas* along with its message length representation. Compared to the case where the topology of the mesh was forced to be regular [Fig. 5(b)], allowing the mesh nodes to be placed strategically decreases the message length further by 10%. Note that the amount of blur is now nonuniformly distributed, occurring mainly in areas with large intersubject anatomical variability as these areas are most susceptible to model overfitting.

Finally, Fig. 5(d) depicts the *content-adaptive, optimally-deformable atlas* along with its message representation. In contrast to the previous models, in which no deformations were allowed, the positions of the mesh nodes in each of the training images differ from their reference positions, and therefore need to be explicitly encoded as well. The variation in length of the three message blocks for increasing values of β, starting at $\beta = 10$, is shown in Fig. 6(b). As expected, the number of bits needed to encode the label probabilities is independent of the deformation flexibility. However, the data message block decreases and the mesh node position block increases for increasingly flexible deformation field models: knowledge about the anatomical labels in the individual training images is effectively transferred into

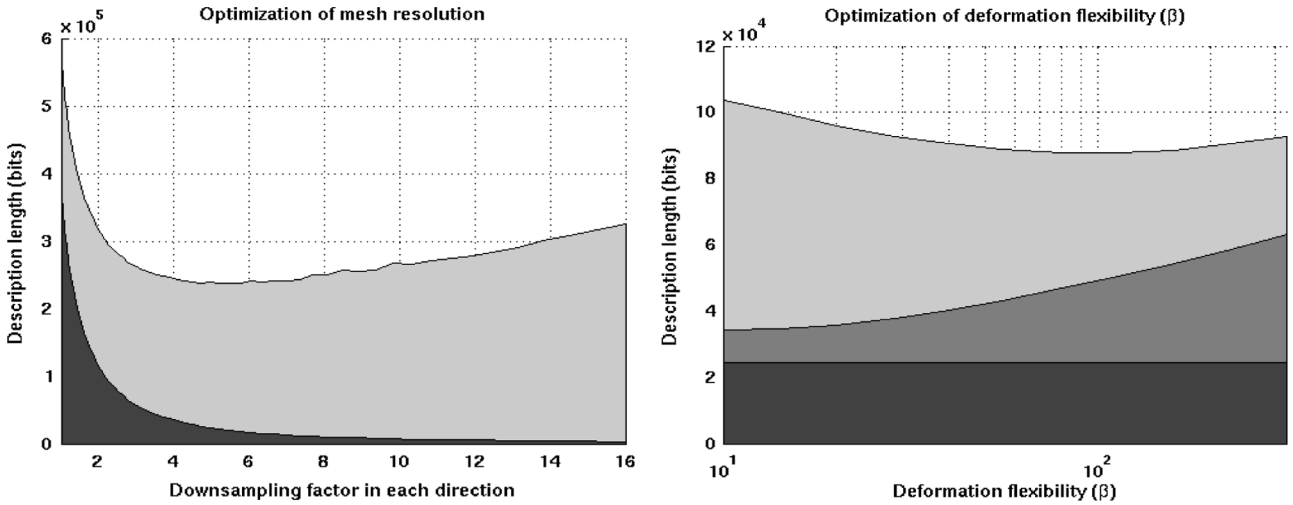

Fig. 6. Optimization of the mesh resolution in regular, nondeformable meshes (a) and the parameter controlling the deformation flexibility of content-adaptive atlases (b) for the first training dataset. The overall message length encoding the training data, as well as the lengths of the constituent message blocks, changes when the parameter of interest varies; the optimum is found at the shortest overall message length. The message blocks are depicted using the same color scheme as in Fig. 5.

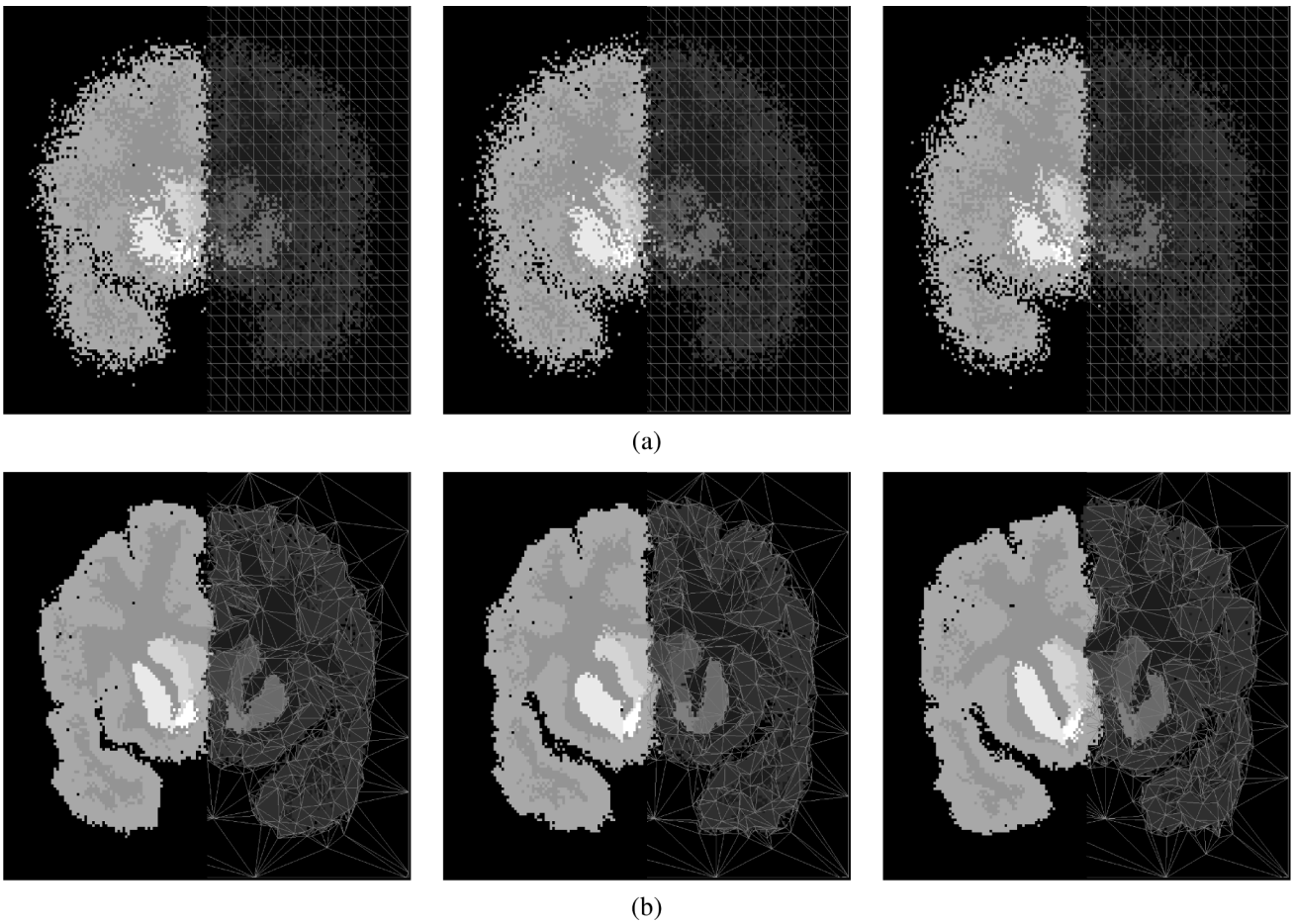

Fig. 7. Samples synthesized using the *optimal-resolution, nondeformable atlas* (a) and the *content-adaptive, optimally-deformable atlas* (b) trained on the first dataset. The right half of the underlying mesh is overlaid on top of the samples in order to visualize the applied deformations. A supplementary animation of more samples is available at http://ieeexplore.ieee.org.

the deformation fields as the training data are better aligned and the atlas gets sharper.

To conclude, we show in Fig. 7(a) and (b) some samples synthesized from the *optimal-resolution, nondeformable atlas*

model and the *content-adaptive, optimally-deformable atlas* model, respectively. It is obvious that the latter model has indeed captured the characteristics of the training data better, explaining its higher compression rates.

VI. Application in 3-D

We here present experiments of the proposed atlas construction technique in 3-D, and show the resulting atlases' potential in fully-automated, pulse sequence-adaptive segmentation of 36 neuroanatomical structures in brain MRI scans. Additional results of the proposed techniques can be found in a recent paper on automated segmentation of the subfields of the hippocampus in ultra-high resolution MRI scans [40].

A. Atlas Construction in 3-D

Fig. 8 shows the *content-adaptive, optimally-deformable atlas* constructed from 3-D manual annotations in four randomly chosen subjects of the IBSR dataset. In these images, each voxel in the entire brain is labeled as one of 36 neuroanatomical structures, including left and right caudate, putamen, pallidum, thalamus, lateral ventricles, hippocampus, amygdala, cerebral, and cerebellar white matter and cortex, and the brain stem. Prior to the atlas computation, the images were coregistered and resampled using the procedure described in Section IV-A, resulting in images of size $177 \times 164 \times 128$.

The employed atlas construction procedure was entirely analogous to the one used in the 2-D case, but using tetrahedral rather than triangular meshes and with the following computational speedups.

- The distance between the nodes in the high-resolution mesh from which the edge collapse operations start, was three times the distance between the voxel centers.
- In the multiresolution approach used to obtain the high-resolution mesh, tethrahedra covering only background were not further subdivided, resulting in less edges to collapse later on.
- The mesh collapse operations were based only on evaluating the sum of the label probabilities message length and the data message length, since the node position message length was observed to have negligible impact on the mesh simplification, and its calculation in 3-D was rather slow in our implementation.
- The code was multithreaded, using multiple processors simultaneously.

It took 34 h to compute the atlas shown in Fig. 8 on a machine with two dual-core Intel Xeon 5140 processors. The computational burden scales essentially linearly with the number of subjects in the training set: computing an atlas from 10 subjects on the same machine increased the computation time to 101 h.

B. Sequence-Adaptive Whole Brain Segmentation

As an example of the potential usage of the proposed atlas models, we here describe a Bayesian method for sequence-adaptive segmentation of 36 brain structures using the tetrahedral atlas of Fig. 8. Building on our earlier work [41], [1], we supplement the *prior* distribution provided by the atlas, which models the generation of images where each voxel is assigned a unique neuroanatomical label, with a *likelihood* distribution that predicts how such label images translate into MRI images, where each voxel has an intensity. Together these distributions form a complete computational model of MRI image formation that we use to obtain fully automated segmentations. While the method described here only segments uni-spectral images, extending it to handle multispectral data is straightforward [41].

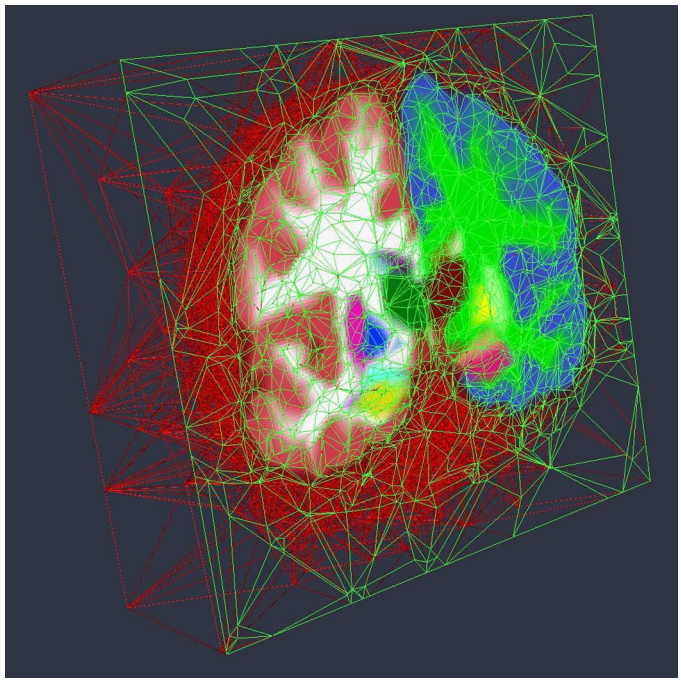

Fig. 8. Optimal tetrahedral mesh-based atlas in 3-D, built from manual annotations of 36 neuroanatomical structures in four subjects. The prior probabilities for the different structures have been color-coded for visualization purposes. The edges of the tetrahedra are shown in red, and the intersections of the faces of the tetrahedra with a cutting plane used for visualization in green.

1) Prior: Once the optimal atlas model and its parameters have been learned from manually annotated training data, the probability of seeing a label image L is given by $p(L \,|\, \hat{\boldsymbol{\alpha}}, \boldsymbol{x}, \hat{\mathcal{K}}) = \prod_{i=1}^{I} p_i(l_i \,|\, \hat{\boldsymbol{\alpha}}, \boldsymbol{x}, \hat{\mathcal{K}})$ (3), where the position of the mesh nodes $\boldsymbol{x}$ is governed by $p(\boldsymbol{x} \,|\, \hat{\beta}, \hat{\boldsymbol{x}}^r, \hat{\mathcal{K}})$ (1). To simplify notation, we will drop the explicit dependency on the learned $\hat{\boldsymbol{\alpha}}, \hat{\beta}, \hat{\boldsymbol{x}}^r$, and $\hat{\mathcal{K}}$ in the remainder, and simply write $p(L \,|\, \boldsymbol{x}) = \prod_{i=1}^{I} p_i(l_i \,|\, \boldsymbol{x})$ and $p(\boldsymbol{x})$ instead.

2) Likelihood: For the likelihood distribution, we employ a model according to which a Gaussian distribution with mean μ_k and variance σ_k^2 is associated with each label k. In order to account for the smoothly varying intensity inhomogeneities or "bias fields" that typically corrupt MR images, we also explicitly model the bias field as a linear combination $\sum_{p=1}^{P} c_p \Psi_p(\cdot)$ of P polynomial basis functions $\Psi_p(\cdot)$. Given label image L, an intensity image $Y = \{y_i, i = 1, \ldots, I\}$ is generated by first drawing the intensity in each voxel independently from the

Gaussian distribution associated with its label, and then adding[4] the local bias field value

$$p(Y \mid L, \boldsymbol{\theta}) = \prod_{i=1}^{I} p_i\left(y_i \mid \mu_{l_i}, \sigma_{l_i}^2, \{c_p\}\right) = \prod_{i=1}^{I} \frac{1}{\sqrt{2\pi\sigma_{l_i}^2}} \exp\left(-\frac{\left(y_i - \mu_{l_i} - \sum_{p=1}^{P} c_p \Psi_p(\boldsymbol{x}_i)\right)^2}{2\sigma_{l_i}^2}\right).$$

Here, the likelihood distribution parameters $\boldsymbol{\theta} = \{\{\mu_k\}, \{\sigma_k^2\}, \{c_p\}\}$ are the means and variances of the Gaussian distributions, as well as the parameters of the bias field model. To complete the model, we specify a uniform prior distribution on these parameters

$$p(\boldsymbol{\theta}) \propto 1.$$

3) Model Parameter Estimation: With the complete generative model in place, Bayesian image segmentation can proceed by first assessing the parameter values $\{\hat{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}\}$ that are most probable in light of the data. We maximize

$$p(\boldsymbol{x}, \boldsymbol{\theta} \mid Y) \propto p(Y \mid \boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{x}) p(\boldsymbol{\theta}) \propto \left(\prod_{i=1}^{I} \sum_{k=1}^{K} p_i(y_i \mid \mu_k, \sigma_k^2, \{c_p\}) p_i(k \mid \boldsymbol{x})\right) p(\boldsymbol{x})$$

which is equivalent to minimizing

$$\sum_{i=1}^{I} \left(-\log\left[\sum_{k=1}^{K} p_i(y_i \mid \mu_k, \sigma_k^2, \{c_p\}) p_i(k \mid \boldsymbol{x})\right]\right) - \log p(\boldsymbol{x}) \quad (10)$$

using a generalized expectation-maximization (GEM) algorithm [31]. We repeatedly calculate a statistical classification that associates each voxel with each of the neuroanatomical labels

$$\Omega_i^k = \frac{p_i(y_i \mid \mu_k, \sigma_k^2, \{c_p\}) p_i(k \mid \boldsymbol{x})}{\sum_{k'} p_i(y_i \mid \mu_{k'}, \sigma_{k'}^2, \{c_p\}) p_i(k' \mid \boldsymbol{x})}$$

and subsequently use this classification to construct an upper bound to (10) that touches it at the current parameter estimates

$$\sum_{i=1}^{I} \left(-\log\left[\prod_{k=1}^{K} \left(\frac{p_i(y_i \mid \mu_k, \sigma_k^2, \{c_p\}) p_i(k \mid \boldsymbol{x})}{\Omega_i^k}\right)^{\Omega_i^k}\right]\right) - \log p(\boldsymbol{x}). \quad (11)$$

For a given position of the mesh nodes $\boldsymbol{x}$, we previously derived closed-form updates for the likelihood distribution parameters $\boldsymbol{\theta}$ that either improve the upper bound—and thus the objective function—or leave it unchanged [41]. After updating $\boldsymbol{\theta}$ this way, the classification and the corresponding upper bound are recalculated, and the estimation of $\boldsymbol{\theta}$ is repeated, until convergence. We then recalculate the upper bound, and optimize it with respect to the mesh node positions $\boldsymbol{x}$, keeping $\boldsymbol{\theta}$ fixed. Optimizing $\boldsymbol{x}$ is a registration process that deforms the atlas mesh towards the current classification, similar to the schemes proposed in [7], [8]. We perform this registration by gradient descent, using the fact that the gradient of (11) with respect to $\boldsymbol{x}$ is given in analytical form. Subsequently, we repeat the optimization of $\boldsymbol{\theta}$ and $\boldsymbol{x}$, each in turn, until convergence.

4) Image Segmentation: Once we have an estimate of the optimal model parameters $\{\hat{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}\}$, we use it to assess the most probable anatomical labeling. Approximating $p(L \mid Y) = \int_{\boldsymbol{x}} \int_{\boldsymbol{\theta}} p(L \mid Y, \boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{x}, \boldsymbol{\theta} \mid Y) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{\theta}$ by $p(L \mid Y, \hat{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}) \propto p(Y \mid L, \hat{\boldsymbol{\theta}}) p(L \mid \hat{\boldsymbol{x}})$, we have

$$\hat{L} = \arg\max_{L} p(L \mid Y) \simeq \arg \max_{\{l_i, i=1,\ldots,I\}} \prod_{i}^{I} p_i\left(y_i \mid \hat{\mu}_{l_i}, \hat{\sigma}_{l_i}^2, \{\hat{c}_p\}\right) p_i(l_i \mid \hat{\boldsymbol{x}})$$

which is obtained by assigning each voxel to the label with the highest posterior probability, i.e., $\hat{l}_i = \arg\max_k \Omega_i^k$.

5) Results: Fig. 9 shows the segmentation results of the proposed algorithm on a high-resolution T1- and T2-weighted scan of two different subjects. The method required no other preprocessing than affinely coregistering the atlas with each image under study [34], and took 25 min computation time on a Intel Core 2 T7600 processor for each subject. Note that the method does not make any assumptions about the MRI scanning protocol used to acquire the images, and is able to automatically adapt to the tissue contrast at hand. While this type of sequence-adaptiveness is now well-established in methods aiming at segmenting the major brain tissue classes [1], [4], [42], [43], this is not the case in state-of-the-art methods for segmenting brain *substructures*, which typically require that all images to be segmented are acquired with a specific image acquisition protocol [7], [26], [44], [45].

A detailed validation of the whole brain segmentation technique proposed here, evaluating the effects on segmentation accuracy of the used pulse sequence(s) and the number of subjects included in the atlas, is outside the scope of this paper and will be published elsewhere.

VII. Discussion and Outlook

In this paper, we addressed the problem of creating probabilistic brain atlases from manually labeled training data. We formulated the atlas construction problem in a Bayesian context, generalizing the generative model implicitly underlying standard average atlases, and comparing alternative models in order to select better representations. We demonstrated, using 2-D training datasets, that this allows us to obtain models that are better at capturing the structure in the training data than conventional probabilistic atlases. We also illustrated, in 3-D, the resulting atlases' potential in sequence-adaptive segmentation of a multitude of brain substructures.

A. Connection to Group-Wise Registration

We described three levels of inference for atlas computation. At the first level, we use an *a priori* specified model, and infer what values its model parameters may take, given the training data. This naturally leads to a so-called *group-wise* registration

[4]Since MR field inhomogeneities are usually assumed multiplicative, we work with logarithmically transformed intensities, rather than with the original image intensities.

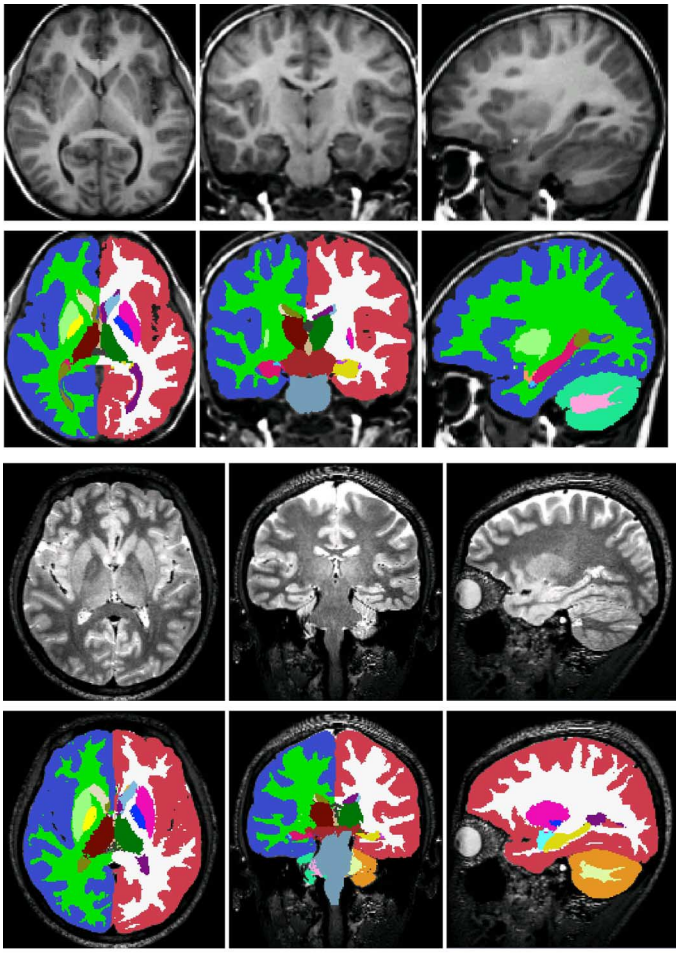

Fig. 9. Application of the atlas of Fig. 8 for sequence-adaptive segmentation of 36 neuroanatomical structures. Results are shown for a high-resolution T1-weighted (top) and a T2-weighted (bottom) brain scan of two different subjects.

algorithm, in which all training datasets are simultaneously aligned with a "average" template that is automatically estimated during the process as well. Similar to existing approaches [16]–[22], the geometry of this average atlas is *unbiased* in that it represents a central tendency among all training datasets, without being affected by the choice of one "representative" dataset in particular.

Our first level of inference differs from other group-wise registration approaches in that the intrinsic coordinate system of the average template is not defined on a regular, high-resolution image grid, as is typically done, but rather on a mesh-based representation in which triangular (or, in 3-D, tetrahedral) elements stretch over a potentially large number of pixels (voxels). This has the implication that an explicit model of the interpolation process is needed, resulting in an iterative algorithm to accurately determine the association of individual pixels with the mesh nodes. In contrast, interpolation can generally be considered a minor issue when dense image grid representations are used, and is typically addressed using simpler, ad hoc schemes.

Our goodness-of-fit criterion, measuring the likelihood of a given set of label probabilities and deformation fields, is closely related to the criterion used in *congealing* approaches [46]. Congealing, when applied to group-wise registration [14], [16], assesses how well a set of spatial transformations aligns a group of images by calculating the sum of voxel-wise entropies. Disregarding interpolation issues and the variable numbers of voxels associated with each mesh node in our model, it is clear that such a sum of entropies is proportional to our likelihood, evaluated at the optimal label probabilities for a given set of deformations. In other words, congealing essentially amounts to a different optimization strategy, in which the joint search space over label probabilities and deformation fields is collapsed into the lower-dimensional space of deformations only, optimizing the label probabilities out for each set of deformations.

To the best of our knowledge, only two other groups have attempted to construct probabilistic atlases from annotated images while simultaneously aligning these images using deformable registration. De Craene *et al.* [11] employed a generative image model in which differences between label images in a training dataset are explained as a combination of deformations and voxel-wise label errors applied to an underlying label image that is shared across all training images [47]. Lorenzen *et al.* [12] constructed an atlas from probabilistic segmentations of brain MR images by minimizing, in each voxel, the Kullback–Leibler distance of the atlas from the probabilistic segmentations. But the atlases obtained by Lorenzen *et al.* and De Craene *et al.* are a normalized voxel-wise *geometric* mean over the training datasets, whereas standard probabilistic atlases are calculated as the *arithmetic* mean. For this reason, these atlases exhibit overly sharp boundaries between structures, and their usefulness as a probabilistic prior in automated segmentation algorithms is therefore questionable.[5] In contrast, our approach is a true generalization of standard probabilistic atlases, based on a generative image model that directly justifies its interpretation as a segmentation prior.

While our first level of inference, similar to other group-wise registration algorithms, jointly optimizes over the average atlas representation and the deformation fields warping it to each of the training images, this results in atlases that are generally biased, as one reviewer pointed out. Indeed, the correct procedure would be to *integrate* over the possible deformation fields when assessing the optimal label probabilities, rather than to *optimize* over them, but unfortunately this ideal approach is computationally unfeasible.[6]

B. Learning the Deformation Field Model

Our second level of inference assesses the most likely flexibility of the deformation field model, given the training data. Using the equivalence between negative log-likelihood and code-length, we showed how this estimation problem can be approximated as a data compression problem that is intuitive to interpret and computationally feasible.

Assessing the flexibility of our deformation model amounts to learning the high-dimensional probability density function (PDF) governing the deformation fields, a problem that has been approached by others in a markedly different way. In [49]–[52], deformation field PDFs were estimated from a number of example deformation fields assumed available for training; these training deformation fields were obtained using

[5]Interestingly, the Kullback–Leibler distance is not symmetric; if Lorenzen *et al.* had swapped the role of their model parameters and the data in their goodness-of-fit criterion, they would have obtained an arithmetic mean in a most natural manner.

[6]See also [48] for a related observation in a different context.

an automated registration algorithm. This inevitably leads to a "chicken and egg" situation, because the generated training samples depend on the deformation field prior used in the registration process, but estimating this prior is exactly the objective [21]. This problem ultimately arises from the lack of notion of *optimality* for deformations aligning brain scans of different individuals: it is not immediately clear how competing priors should be compared. We are not confronted with this difficulty because, rather than trying to estimate a PDF that describes "true" or physically plausible deformation fields, our goal is to model pixelated, manually-labeled example segmentations, which is objectively quantified by the description length criterion. Note that, in contrast to [49]–[52], we do not train our model directly on a set of deformation fields: the explicit calculation of deformation fields in our approach only arises as a mathematical means to approximate our objective function.

C. Content-Adaptive Mesh Representation

At our third level of inference, we compare competing mesh representations by evaluating how compactly they encode the training data. This allows us to determine the optimal resolution of regular meshes for a given training dataset, automatically avoiding overfitting and ensuring that the resulting atlases are sufficiently blurry to generalize to unseen cases. It also allows us to construct *content-adaptive* meshes, in which certain areas have a much higher density of mesh nodes than others. When this is combined with a deformation model, large areas that cover the same anatomical label and that exhibit relatively smooth boundaries, such as the ventricles and deep gray matter structures, can be fully represented by a set of mesh nodes located along their boundaries [see Fig. 5(d)]. In this sense, our representation can be related to *statistical shape models* (for instance, [53]–[55]), in which objects are described by their boundaries alone.[7] In contrast to such shape models, however, our approach also allows areas in which shape characteristics fluctuate widely between individuals, such as cortical areas, to be encoded by a "blurry" probabilistic representation, rather than by explicitly describing the convoluted details of each individual's label boundaries. Which of these two representations is most advantageous in each brain area is automatically determined by comparing their respective code lengths during the mesh simplification procedure.

While the construction of our content-adaptive meshes may seem prohibitively time consuming, especially when the technique is applied in 3-D, we do not consider this a liability. Manually outlining dozens of structures in volumetric brain scans is notoriously time consuming, requiring up to one week for a single scan [26]. In this light, thoroughly analyzing the resulting training data using ubiquitous and increasingly powerful computing hardware is unlikely to be a bottleneck. Furthermore, computation time spent constructing sparse atlas representations, which only needs to be done once for a given training dataset, can significantly save computation time in segmentation algorithms warping the resulting atlases, potentially benefitting the analysis of thousands of images. Reducing the dimensionality of warping problems by eliminated superfluous degrees of freedom, while only an accidental by-product of our atlas encoding approach, is a valuable goal in itself in medical image registration [56], [57].

[7]Although our deformation model, with its single parameter, is obviously no match for sophisticated shape models; see also later.

D. Difficulties in Validation

The goal of our atlas construction work is to provide automated brain MRI segmentation techniques with priors that encode the normal anatomical variability in the population under study as accurately as possible. As such, the ultimate test of the atlases we generate would be to check how well they predict brain delineations in subjects not included in the training database. A typical way to do this would be so-called *leave-one-out cross-validation*: a single subject is removed from the training set, an atlas is computed from the remaining subjects, and the probability with which the resulting atlas predicts the left-out subject is evaluated; this process is then repeated for all subjects, and the results are averaged.

Unfortunately, we have not been able to perform such a cross-validation because of practical difficulties. A first obstacle arises from the fact that our atlases are deformable: evaluating the probability of observing a given label image involves integrating over all possible atlas deformations. In theory, such a problem can be numerically approximated using Monte Carlo techniques,[8] by drawing enough samples from the deformation field prior, and averaging the probabilities of observing the data under each of the deformations. In practice, however, such an approach does not provide useful answers in a high-dimensional problem as ours, as none of the samples that can be generated in a practical amount of time will provide a reasonable alignment to the validation data. We, therefore, experimented with annealed importance sampling (AIS) [58], a technique that addresses this issue by building a smooth transition path of intermediate distributions between the prior and the posterior, and collating the results of sampling measurements collected while proceeding through the chain of distributions. In our implementation, the intermediate distributions were obtained by applying a varying degree of spatial smoothing to our atlases. Unfortunately, we were not able to obtain sound results with this technique as the contributions of the distributions close to the posterior proved to be especially hard to estimate reliably.

A second problem is that manual annotations typically contain a number of small spots with a different label than the one expected at the corresponding location, such as isolated strands of background label in the middle of the brain (see Fig. 4). Such spots preclude a validation of even our procedure to assess the optimal resolution of regular meshes, as their probability of occurrence is technically zero under atlases constructed over a wide range of mesh resolutions. The underlying problem is that we use the *optimal* values of the label probability parameters after training only, whereas a full Bayesian treatment would use the ensemble of *all* label probability values, each weighed by its posterior under the training data. Such a model would not assign a zero probability to validation datasets, but deriving it in practice is greatly complicated by our interpolation model and by our prior, and we have therefore not pursued this option further.

[8]Of course, deterministic approximations such as the Laplace approximation used earlier are also possible, but evaluating these approximations is exactly part of the objective here.

E. Outlook and Future Work

The generalized probabilistic atlas model proposed in this paper is not beyond improvement. In particular, our deformation field model has only one single parameter: the deformation field flexibility. While such nonspecific models are the norm in the field of nonrigid registration and lie at the heart of some of the most advanced brain MRI segmentation techniques available to date [6], more powerful deformation models can be constructed if they have more parameters to be trained [49]–[52]. More extensive parametrizations, regulating different aspects of deformation in individual triangles (or, in 3-D, tetrahedra) or in triangles sharing the same anatomical labels, is an option we plan to explore in the future. Provided that the bits needed to encode them are not ignored (as we have done here for our flexibility parameter), the appropriateness of adding such extra parameters can be directly evaluated using the code length criterion. More generally, alternative models of deformation, such as those based on the large deformation or diffeomorphic framework [59], [60], as opposed to the small deformation setting [61] used here, can be tried and compared by evaluating their respective code lengths.

While this paper concentrated on building priors from manually labeled training data, we also demonstrated the potential usage of the resulting atlases in sequence-adaptive segmentation of dozens of neuroanatomical structures in MRI scans of the head. Further developing the proposed technique and carefully evaluating its segmentation performance will be the focus of our future research.

APPENDIX A
DEFORMATION FIELD PENALTY

The prior proposed by Ashburner *et al.* [21] is defined as follows in 2-D (the 3-D case is analog). Let $\boldsymbol{x}^r_{t,j} = (u^r_{t,j}, v^r_{t,j})$ and $\boldsymbol{x}_{t,j} = (u_{t,j}, v_{t,j})$ denote the position of the jth corner of triangle t in the mesh at reference position and after deformation, respectively. The affine mapping of the triangle $\mathbf{M}$ is then obtain by

$$\mathbf{M} = \begin{bmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} u_{t,1} & u_{t,2} & u_{t,3} \\ v_{t,1} & v_{t,2} & v_{t,3} \\ 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} u^r_{t,1} & u^r_{t,2} & u^r_{t,3} \\ v^r_{t,1} & v^r_{t,2} & v^r_{t,3} \\ 1 & 1 & 1 \end{bmatrix}^{-1}.$$

The Jacobian matrix $\mathbf{J}$ of this mapping, given by

$$\mathbf{J} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}$$

can be decomposed into two rotations $\mathbf{U}$ and $\mathbf{V}$ and a diagonal matrix $\mathbf{S}$, using a singular value decomposition (SVD): $\mathbf{J} = \mathbf{U}\mathbf{S}\mathbf{V}^T$, where

$$\mathbf{S} = \begin{bmatrix} s_1 & 0 \\ 0 & s_2 \end{bmatrix}.$$

Ashburner's penalty for each triangle is based on its area and on the singular values s_1 and s_2, which represent relative stretching in orthogonal directions:

$$U^{\mathcal{K}}_t(\boldsymbol{x} \,|\, \boldsymbol{x}^r) = A^r_t \cdot \left(1 + \prod_{i=1}^{2} s_i\right) \cdot \sum_{i=1}^{2} \left(s_i^2 + 1/s_i^2 - 2\right)$$

where A^r_t denotes the area of the triangle in the reference position. This can be conveniently calculated without performing a SVD as

$$U^{\mathcal{K}}_t(\boldsymbol{x} \,|\, \boldsymbol{x}^r) = A^r_t \cdot (1 + |J|) \cdot \left(||J||_2^2 \cdot (1 + 1/|J|^2) - 4\right).$$

APPENDIX B
INTEGRATING OVER $\boldsymbol{x}$

We here derive the approximation used to obtain (8). Assuming that $p(L_m \,|\, \boldsymbol{\alpha}, \boldsymbol{x}^m, \mathcal{K}) p(\boldsymbol{x}^m \,|\, \beta, \boldsymbol{x}^r, \mathcal{K})$ has a peak at a position $\boldsymbol{x}^m_\alpha$, we may approximate $p(L_m \,|\, \boldsymbol{\alpha}, \beta, \boldsymbol{x}^r, \mathcal{K})$ using Laplace's method, i.e., by locally approximating the integrand by an unnormalized Gaussian

$$\begin{aligned} &p(L_m \,|\, \boldsymbol{\alpha}, \beta, \boldsymbol{x}^r, \mathcal{K}) \\ &\quad = \int_{\boldsymbol{x}^m} p(L_m \,|\, \boldsymbol{\alpha}, \boldsymbol{x}^m, \mathcal{K}) p(\boldsymbol{x}^m \,|\, \beta, \boldsymbol{x}^r, \mathcal{K}) \mathrm{d}\boldsymbol{x}^m \\ &\quad \simeq \int_{\boldsymbol{x}^m} p(L_m \,|\, \boldsymbol{\alpha}, \boldsymbol{x}^m_\alpha, \mathcal{K}) p\left(\boldsymbol{x}^m_\alpha \,|\, \beta, \boldsymbol{x}^r, \mathcal{K}\right) \\ &\qquad \times \exp\left(-\frac{1}{2} \left(\boldsymbol{x}^m - \boldsymbol{x}^m_\alpha\right)^T \boldsymbol{I}^m \left(\boldsymbol{x}^m - \boldsymbol{x}^m_\alpha\right)\right) \mathrm{d}\boldsymbol{x}^m \\ &\quad = p\left(L_m \,|\, \boldsymbol{\alpha}, \boldsymbol{x}^m_\alpha, \mathcal{K}\right) \cdot p\left(\boldsymbol{x}^m_\alpha \,|\, \beta, \boldsymbol{x}^r, \mathcal{K}\right) \\ &\qquad \cdot \sqrt{\frac{(2\pi)^{2N}}{\det(\boldsymbol{I}^m)}} \end{aligned} \tag{12}$$

where

$$\boldsymbol{I}^m = D^2_{\boldsymbol{x}}[-\log p(L_m \,|\, \boldsymbol{\alpha}, \boldsymbol{x}, \mathcal{K}) - \log p(\boldsymbol{x} \,|\, \beta, \boldsymbol{x}^r, \mathcal{K})]|_{\boldsymbol{x}=\boldsymbol{x}^m_\alpha}.$$

Letting $[\boldsymbol{x}^m_\alpha]_n$ denote the position of the n-th node in $\boldsymbol{x}^m_\alpha$, and $\boldsymbol{x}^m_{\alpha \setminus n}$ the position of all other nodes, we may assess the prior $p(\boldsymbol{x}^m_\alpha \,|\, \beta, \boldsymbol{x}^r, \mathcal{K})$ using the pseudo-likelihood approximation [32]:

$$\begin{aligned} &p\left(\boldsymbol{x}^m_\alpha \,|\, \beta, \boldsymbol{x}^r, \mathcal{K}\right) \\ &\quad \simeq \prod_{n=1}^{N} p\left([\boldsymbol{x}^m_\alpha]_n \,|\, \boldsymbol{x}^m_{\alpha \setminus n}, \beta, \boldsymbol{x}^r, \mathcal{K}\right) \\ &\quad = \prod_{n=1}^{N} \frac{\exp\left(-\frac{U(\boldsymbol{x}^m_\alpha \,|\, \boldsymbol{x}^r, \mathcal{K})}{\beta}\right)}{\int_{[\boldsymbol{x}^m_\alpha]_n} \exp\left(-\frac{U(\boldsymbol{x}^m_\alpha \,|\, \boldsymbol{x}^r, \mathcal{K})}{\beta}\right) \mathrm{d}\,[\boldsymbol{x}^m_\alpha]_n} \\ &\quad \simeq \prod_{n=1}^{N} \frac{\exp\left(-\frac{U(\boldsymbol{x}^m_\alpha \,|\, \boldsymbol{x}^r, \mathcal{K})}{\beta}\right)}{\exp\left(-\frac{U(\boldsymbol{x}^{m \,|\, n}_\alpha \,|\, \boldsymbol{x}^r, \mathcal{K})}{\beta}\right) \sqrt{(2\pi)^2 / \det(\boldsymbol{J}^m_n)}} \end{aligned} \tag{13}$$

where a local Laplace approximation was used in the last step, and the notations $\boldsymbol{x}^{m \,|\, n}_\alpha$ and $\boldsymbol{J}^m_n$ are as defined in Section III-B.

Finally, ignoring interdependencies between neighboring mesh nodes in $\boldsymbol{I}^m$, and using the notation $\boldsymbol{I}_n^m$ as defined in Section III-B, we have

$$\det(\boldsymbol{I}^m) \simeq \prod_{n=1}^{N} \det\left(\boldsymbol{I}_n^m\right). \tag{14}$$

Plugging (13) and (14) into (12), we obtain (8).

Appendix C
Integrating Over $\boldsymbol{\alpha}$

We here derive the approximation to

$$\int_{\boldsymbol{\alpha}} \left(\prod_{m=1}^{M} p(L_m \,|\, \boldsymbol{\alpha}, \hat{\boldsymbol{x}}^m, \mathcal{K}) \right) p(\boldsymbol{\alpha}) \mathrm{d}\boldsymbol{\alpha}$$

used to obtain (9). Substituting $p(L_m \,|\, \boldsymbol{\alpha}, \hat{\boldsymbol{x}}^m, \mathcal{K})$ with the lower bound

$$\prod_{i=1}^{I} \prod_{n=1}^{N} \left(\frac{\alpha_n^{l_i^m} \, \hat{\phi}_n^m(\boldsymbol{x}_i)}{\hat{W}_{i,n}^m} \right)^{\hat{W}_{i,n}^m}$$

factorizes the integrand over the mesh nodes

$$\int_{\boldsymbol{\alpha}} \left(\prod_{m=1}^{M} p(L_m \,|\, \boldsymbol{\alpha}, \hat{\boldsymbol{x}}^m, \mathcal{K}) \right) p(\boldsymbol{\alpha}) \mathrm{d}\boldsymbol{\alpha} \simeq \prod_{n=1}^{N} \left(\prod_{m=1}^{M} \prod_{i=1}^{I} \left(\frac{\hat{\phi}_n^m(\boldsymbol{x}_i)}{\hat{W}_{i,n}^m} \right)^{\hat{W}_{i,n}^m} \right) \cdot I_n \tag{15}$$

where

$$I_n = \int_{\boldsymbol{\alpha}_n} \left(\prod_{k=1}^{K} \left(\alpha_n^k\right)^{\hat{\alpha}_n^k \cdot \hat{N}_n} \right) p(\boldsymbol{\alpha}_n) \mathrm{d}\boldsymbol{\alpha}_n$$

with $\hat{N}_n = \sum_{m=1}^{M} \sum_{i=1}^{I} \hat{W}_{i,n}^m$. For the case $K = 3$, the prior $p(\boldsymbol{\alpha}_n)$ is flat and I_n is given by

$$\begin{aligned} & 2! \cdot \frac{\prod_{k=1}^{K=3} \Gamma\left(\hat{\alpha}_n^k \cdot \hat{N}_n + 1\right)}{\Gamma(\hat{N}_n + 3)} \\ &= \frac{2!\Gamma(\hat{N}_n + 1)}{\Gamma(\hat{N}_n + 3)} \cdot \frac{\prod_{k=1}^{K=3} \Gamma\left(\hat{\alpha}_n^k \cdot \hat{N}_n + 1\right)}{\Gamma(\hat{N}_n + 1)} \\ &\simeq \frac{2!\Gamma(\hat{N}_n + 1)}{\Gamma(\hat{N}_n + 3)} \cdot \prod_{k=1}^{K=3} \left(\hat{\alpha}_n^k\right)^{\hat{\alpha}_n^k \cdot \hat{N}_n} \\ &= \frac{2!\Gamma(\hat{N}_n + 1)}{\Gamma(\hat{N}_n + 3)} \cdot \prod_{m=1}^{M} \prod_{i=1}^{I} \left(\hat{\alpha}_n^{l_i^m}\right)^{\hat{W}_{i,n}^m} \end{aligned}$$

where the next-to-last step is based on Stirling's approximation for the Gamma function $\Gamma(x+1) \simeq x^x e^{-x}$, and on the fact that $\sum_k^K \hat{\alpha}_n^k = 1$. In the general case $K \geq 3$, the prior $p(\boldsymbol{\alpha}_n)$ only allows nonzero probabilities for three labels simultaneously but is otherwise flat, so that we have

$$I_n \simeq \underbrace{\binom{K}{3} \cdot \frac{2!\Gamma(\hat{N}_n + 1)}{\Gamma(\hat{N}_n + 3)}}_{\hat{R}_n} \cdot \prod_{m=1}^{M} \prod_{i=1}^{I} \left(\hat{\alpha}_n^{l_i^m}\right)^{\hat{W}_{i,n}^m}.$$

Plugging this result into (15) and rearranging factors, we finally obtain

$$\begin{aligned} & \int_{\boldsymbol{\alpha}} \left(\prod_{m=1}^{M} p(L_m \,|\, \boldsymbol{\alpha}, \hat{\boldsymbol{x}}^m, \mathcal{K}) \right) p(\boldsymbol{\alpha}) \mathrm{d}\boldsymbol{\alpha} \\ &\qquad \simeq \left(\prod_{m=1}^{M} \prod_{i=1}^{I} \underbrace{\prod_{n=1}^{N} \left(\frac{\hat{\alpha}_n^{l_i^m} \, \hat{\phi}_n^m(\boldsymbol{x}_i)}{\hat{W}_{i,n}^m} \right)^{\hat{W}_{i,n}^m}}_{p_i(l_i^m \,|\, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{x}}^m, \mathcal{K})} \right) \cdot \prod_{n=1}^{N} \hat{R}_n \end{aligned}$$

which explains (9).

Acknowledgment

The author would like to thank B. Fischl for providing the high resolution T2-weighted scan, L. Zöllei for proofreading the manuscript, and the anonymous reviewers for their detailed and constructive comments.

References

[1] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "Automated model-based tissue classification of MR images of the brain," *IEEE Trans. Med. Imag.*, vol. 18, no. 10, pp. 897–908, Oct. 1999.

[2] A. Zijdenbos, R. Forghani, and A. Evans, "Automatic "pipeline" analysis of 3-D MRI data for clinical trials: Application to multiple sclerosis," *IEEE Trans. Med. Imag.*, vol. 21, no. 10, pp. 1280–1291, Oct. 2002.

[3] B. Fischl, D. Salat, A. van der Kouwe, N. Makris, F. Segonne, B. Quinn, and A. Dalea, "Sequence-independent segmentation of magnetic resonance images," *NeuroImage*, vol. 23, pp. S69–S84, 2004.

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

[5] M. Prastawa, J. Gilmore, W. Lin, and G. Gerig, "Automatic segmentation of MR images of the developing newborn brain," *Med. Image Anal.*, vol. 9, pp. 457–466, 2005.

[6] R. 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, Oct. 2006.

[7] K. Pohl, J. Fisher, W. Grimson, R. Kikinis, and W. Wells, "A Bayesian model for joint segmentation and registration," *NeuroImage*, vol. 31, no. 1, pp. 228–239, May 2006.

[8] E. D'Agostino, F. Maes, D. Vandermeulen, and P. Suetens, "A unified framework for atlas based brain image segmentation and registration," in *WBIR 2006*. New York: Springer-Verlag, 2006, vol. 4057, Lecture Note Computer Science, pp. 136–143.

[9] S. Awate, T. Tasdizen, N. Foster, and R. Whitaker, "Adaptive Markov modeling for mutual-information-based, unsupervised MRI brain-tissue classification," *Med. Image Anal.*, vol. 10, no. 5, pp. 726–739, 2006.

[10] J. Ashburner, "Computational neuroanatomy," Ph.D. dissertation, Univ. London, London, U.K., 2000.

[11] M. De Craene, A. du Bois d'Aische, B. Macq, and S. Warfield, "Multi-subject registration for unbiased statistical atlas construction," in *MICCAI 2004*. New York: Springer-Verlag, 2004, vol. 3216, Lecture Notes Computer Science, pp. 655–662.

[12] P. Lorenzen, M. Prastawa, B. Davis, G. Gerig, E. Bullitt, and S. Joshi, "Multi-modal image set registration and atlas formation," *Med. Image Anal.*, vol. 10, pp. 440–451, 2006.

[13] B. Avants and J. Gee, "Geodesic estimation for large deformation anatomical shape averaging and interpolation," *NeuroImage*, vol. 23, no. 1, pp. 139–150, 2004.

[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 *MICCAI 2001*. New York: Springer-Verlag, 2001, vol. 2208, Lecture Notes Computer Science, pp. 266–274.

[15] D. Seghers, E. D'Agostino, F. Maes, D. Vandermeulen, and P. Suetens, "Construction of a brain template from MR images using state-of-the-art registration and segmentation techniques," in *MICCAI 2004*. New York: Springer-Verlag, 2004, vol. 3216, Lecture Notes Computer Science, pp. 696–703.

[16] L. Zöllei, "A unified information theoretic framework for pair- and group-wise registration of medical images," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, 2006.

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

[18] C. Studholme, "Simultaneous population based image alignment for template free spatial normalisation of brain anatomy," *Biomed. Image Registration—WBIR*, vol. 2717, pp. 81–90, 2003.

[19] S. Joshi, B. Davis, M. Jomier, and G. Gerig, "Unbiased diffeomorphic atlas construction for computational anatomy," *NeuroImage*, vol. 23, no. 1, pp. 151–160, 2004.

[20] J. Ashburner, J. Andersson, and K. Friston, "High-dimensional image registration using symmetric priors," *NeuroImage*, vol. 9, no. 6, pt. 1, pp. 619–628, Jun. 1999.

[21] J. Ashburner, J. Andersson, and K. Friston, "Image registration using a symmetric prior-in three dimensions," *Human Brain Mapp.*, vol. 9, no. 4, pp. 212–225, Apr. 2000.

[22] C. 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," *Inf. Process. Med. Imag.*, vol. 19, pp. 1–14, 2005.

[23] D. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[24] J. Marroquin, B. V. S. Botello, F. Calderon, and A. Fernandez-Bouzas, "An accurate and efficient bayesian method for automatic segmentation of brain MRI," *IEEE Trans. Med. Imag.*, vol. 21, no. 8, pp. 934–945, Aug. 2002.

[25] C. Cocosco, A. Zijdenbos, and A. Evans, "A fully automatic and robust brain MRI tissue classification method," *Med. Image Anal.*, vol. 7, pp. 513–527, 2003.

[26] 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, pp. 341–355, Jan. 2002.

[27] K. Van Leemput, "Probabilistic brain atlas encoding using Bayesian inference," in *MICCAI 2006*. New York: Springer-Verlag, 2006, vol. 4190, Lecture Notes Computer Science, pp. 704–711.

[28] J. Munkres, *Elements of Algebraic Topology*. New York: Perseus, 1984, ch. 1, p. 7.

[29] G. Christensen, "Deformable shape models for anatomy," Ph.D. dissertation, Washington Univ., Saint Louis, MO, Aug. 1994.

[30] M. Nielsen, P. Johansen, A. Jackson, and B. Lautrup, "Brownian warps: A least committed prior for non-rigid registration," in *MICCAI 2002*. New York: Springer-Verlag, 2002, vol. 2489, Lecture Notes Computer Science, pp. 557–564.

[31] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc.*, vol. 39, pp. 1–38, 1977.

[32] J. Besag, "Statistical analysis of non-lattice data," *Statistician*, vol. 24, no. 3, pp. 179–195, 1975.

[33] Internet brain segmentation repository (IBSR) [Online]. Available: http://www.cma.mgh.harvard.edu/ibsr

[34] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Trans. Med. Imag.*, vol. 16, no. 2, pp. 187–198, Apr. 1997.

[35] H. Hoppe, T. DeRose, T. Duchamp, J. McDonald, and W. Stuetzle, "Mesh optimization," in *Proc. 20th Annu. Conf. Comput. Graphics Interactive Techniques*, 1993, pp. 19–26.

[36] H. Hoppe, "Progressive meshes," in *ACM SIGGRAPH 1996*, 1996, pp. 99–108.

[37] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes in C. The Art of Scientific Computing*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 1992.

[38] J. Besag, "On the statistical analysis of dirty pictures," *J. R. Stat. Soc. Series B (Methodological)*, vol. 48, no. 3, pp. 259–302, 1986.

[39] S. Duane, A. Kennedy, B. Pendleton, and D. Roweth, "Hybrid Monte Carlo," *Phys. Lett. B*, vol. 195, no. 2, pp. 216–222, 1987.

[40] K. Van Leemput, A. Bakkour, T. Benner, G. Wiggins, L. Wald, J. Augustinack, B. Dickerson, P. Golland, and B. Fischl, "Model-based segmentation of hippocampal subfields in ultra-high resolution in vivo MRI," in *MICCAI 2008*. New York: Springer-Verlag, 2008, vol. 5241, Lecture Notes Computer Science, pp. 235–243.

[41] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "Automated model-based bias field correction of MR images of the brain," *IEEE Trans. Med. Imag.*, vol. 18, no. 10, pp. 885–896, Oct. 1999.

[42] D. Pham and J. Prince, "Adaptive fuzzy segmentation of magnetic resonance images," *IEEE Trans. Med. Imag.*, vol. 18, no. 9, pp. 737–752, Sep. 1999.

[43] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE Trans. Med. Imag.*, vol. 20, no. 1, pp. 45–57, 2001.

[44] B. Patenaude, S. Smith, D. Kennedy, and M. Jenkinson, Bayesian shape and appearance models Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB), Univ. Oxford, Oxford, U.K., Tech. Rep. TR07BP1, 2007.

[45] Z. Tu, K. Narr, P. Dollar, I. Dinov, P. Thompson, and A. Toga, "Brain anatomical structure segmentation by hybrid discriminative/generative models," *IEEE Trans. Med. Imag.*, vol. 27, no. 4, pp. 495–508, Apr. 2008.

[46] E. Learned-Miller, "Data driven image models through continuous joint alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 236–250, Feb. 2006.

[47] S. Warfield, K. Zou, and W. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans. Med. Imag.*, vol. 23, no. 7, pp. 903–921, Jul. 2004.

[48] S. Allassonnière, Y. Amit, and A. Trouvé, "Towards a coherent statistical framework for dense deformable template estimation," *J. R. Stat. Soc.: Series B (Stat. Methodol.)*, vol. 69, no. 1, pp. 3–29, 2007.

[49] J. Gee and R. Bajcsy, *Elastic Matching: Continuum Mechanical and Probabilistic Analysis*. New York: Academic, 1999, ch. 11, pp. 183–197.

[50] D. Rueckert, A. Frangi, and J. Schnabel, "Automatic construction of 3-D statistical deformation models of the brain using nonrigid registration," *IEEE Trans. Med. Imag.*, vol. 22, no. 8, pp. 1014–1025, Aug. 2003.

[51] S. Joshi, "Large deformation diffeomorphisms and Gaussian random fields for statistical characterization of brain submanifolds," Ph.D. dissertation, Washington Univ., St. Louis, MO, 1998.

[52] Z. Xue, D. Shen, and C. Davatzikos, "Statistical representation of high-dimensional deformation fields with application to statistically constrained 3D warping," *Med. Image Anal.*, vol. 10, no. 5, pp. 740–751, Oct. 2006.

[53] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active shape models—Their training and application," *Comput. Vis. Image Understand.*, vol. 61, no. 1, pp. 38–59, 1995.

[54] A. Kelemen, G. Székely, and G. Gerig, "Elastic model-based segmentation of 3-D neuroradiological data sets," *IEEE Trans. Med. Imag.*, vol. 18, no. 10, pp. 828–839, Oct. 1999.

[55] S. Pizer, P. Fletcher, S. Joshi, A. Thall, J. Chen, Y. Fridman, D. Fritsch, A. Gash, J. Glotzer, M. Jiroutek, C. Lu, K. Muller, G. Tracton, P. Yushkevich, and E. Chaney, "Deformable M-Reps for 3D medical image segmentation," *Int. J.Comput. Vis.*, vol. 55, no. 2, pp. 85–106, 2003.

[56] S. Timoner, "Compact representations for fast nonrigid registration of medical images," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, 2003.

[57] G. Rohde, A. Aldroubi, and B. Dawant, "The adaptive bases algorithm for intensity-based nonrigid image registration," *IEEE Trans. Med. Imag.*, vol. 22, no. 11, pp. 1470–1479, Nov. 2003.

[58] R. Neal, "Annealed importance sampling," *Statist. Comput.*, vol. 11, no. 2, pp. 125–139, 2001.

[59] G. Christensen, R. Rabbitt, and M. Miller, "Deformable templates using large deformation kinematics," *IEEE Trans. Image Process.*, vol. 5, no. 10, pp. 1435–1447, 1996.

[60] M. Beg, M. Miller, A. Trouvé, and L. Younes, "Computing large deformation metric mappings via geodesic flows of diffeomorphisms," *Int. J. Comput. Vis.*, vol. 61, no. 2, pp. 139–157, 2005.

[61] M. Miller, A. Banerjee, G. Christensen, S. Joshi, N. Khaneja, U. Grenander, and L. Matejic, "Statistical methods in computational anatomy," *Stat. Methods Med. Res.*, vol. 6, no. 3, pp. 267–299, 1997.