Probabilistic Brain Atlas Encoding Using Bayesian Inference

Koen Van Leemput

Helsinki Medical Imaging Center, Helsinki University Central Hospital, Finland

Abstract. This paper addresses the problem of creating probabilistic brain atlases from manually labeled training data. We propose a general mesh-based atlas representation, and compare different atlas models by evaluating their posterior probabilities and the posterior probabilities of their parameters. Using such a Baysian framework, we show that the widely used "average" brain atlases constitute relatively poor priors, partly because they tend to overfit the training data, and partly because they do not allow to align corresponding anatomical features across datasets. We also demonstrate that much more powerful representations can be built using content-adaptive meshes that incorporate non-rigid deformation field models. We believe extracting optimal prior probability distributions from training data is crucial in light of the central role priors play in many automated brain MRI analysis techniques.

1 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 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 in the form of *probabilistic atlases* [1–6]. Typically, such atlases are created by voxel-wise averaging of neuroanatomical labels over a collection of manually labeled training datasets. In such an approach, the training datasets are first registered together, and the prior probability of each voxel being occupied by a particular structure is calculated as the relative frequency that structure occurred at that voxel across the training datasets.

While widely used, the quality of such "average" atlases as prior probability distributions has, to our knowledge, never been thoroughly investigated, and several open issues remain. In [6], for example, the authors are faced with the problem of creating a probabilistic atlas for newborn brain from only three training datasets. It is clear that, due to the enormous variability in cortical patterns across individuals, the average of three segmentations generalizes poorly to subjects not included in the training database, and the authors decided to blur their average atlas. But how much blurring should be used to obtain the "optimal" atlas? Is there still a need for blurring when more training data is used?

Another question relates to the use of non-rigid registration techniques. Often atlases are constructed based on affine co-registrations of the training datasets, but are warped using non-rigid registration during the segmentation phase [7, 5, 2, 4]. While this inconsistency can be overcome by deforming the training data during the atlas construction phase itself [8, 9], a central question remains: how should the flexibility of the deformation models be chosen?

In this paper we explore ways to refine our atlas construction abilities beyond those currently available. From the discussion above, it is clear that finding good atlas models cannot simply be guided by how well a model describes the available training data: more complex models can always fit the training data better, leading to implausible, over-parameterized results. Rather, we will compare different models by evaluating their posterior probabilities and the posterior probabilities of their parameters. It is well-known that complex models are self-penalizing under Bayes' rule; we will show that a rigorous Bayesian approach is able to, among other things, provide quantitative answers to the questions raised above.

For the remainder of this paper, we will work with 2-dimensional image domains, with the understanding that the proposed techniques translate directly into 3 dimensions as well.

2 Mesh-based atlas models

Let $L = \{l_i, i = 1, 2, \ldots, I\}$ be a manually labeled image with a total of I pixels, where $l_i \in \{1, 2, \ldots, K\}$ denotes the one of K possible labels assigned to pixel i. Partitioning the image domain D into T non-overlapping triangular elements, denoted by $D_t, t = 1, 2, \ldots T$, so that $D = \cup_{t=1}^{T} D_t$, we model the probability of having label l_i at pixel i by interpolation from the values at the nodes of the element D_t in which i falls:

$$p(l_i|\boldsymbol{\alpha}_{t,j}, \boldsymbol{x}_{t,j}) = \sum_{j=1}^{3} \alpha_{t,j}^{l_i} \varphi_{t,j}(\boldsymbol{x}_i). \tag{1}$$

In equation 1, $\boldsymbol{x}_i$ is the position of the the ith pixel, and $\boldsymbol{\alpha}_{t,j} = \{\alpha_{t,j}^1, \alpha_{t,j}^2, \ldots, \alpha_{t,j}^K\}$ denotes the set of label probabilities at the jth node of D_t. Furthermore, $\boldsymbol{x}_{t,j}$ is the position of the jth node of D_t, and $\varphi_{t,j}(\boldsymbol{x})$ is the linear interpolation basis function associated with this node.

For notational convenience, we will index the mesh nodes by $n = 1, 2, \ldots, N$ for the remainder of the paper, keeping in mind that each mesh node is typically shared among several triangles. Using this notation, equation 1 can be extended to cover the whole image domain D as follows:

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

where $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_N\}$ and $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ is the set of the label probabilities and the positions of the mesh nodes, respectively, $\mathcal{K}$ denotes a

simplicial complex specifying the mesh connectivity, and $\phi_n(\boldsymbol{x})$ is the sum of the interpolation basis functions $\varphi_{t,j}(\boldsymbol{x})$ of the elements attached to node n.

Finally, assuming conditional independence of the labels between pixels given the mesh parameters, we have $p(L|\boldsymbol{\alpha}, \boldsymbol{x}, \mathcal{K}) = \prod_{i=1}^{I} p(l_i|\boldsymbol{\alpha}, \boldsymbol{x}, \mathcal{K})$ for the probability of seeing label image L.

2.1 First level of inference

Given certain training data in the form of M label images $L_m, m = 1, 2, \ldots, M$, and letting $\{\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^M\}$ denote the positions of the mesh nodes in each of the expert labelings, we may wish to infer what values of $\boldsymbol{\alpha}$ and $\{\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^M\}$ best fit the training data. To this end, we define a topology-preserving Markov random field prior on the position of the mesh nodes:

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

$$\text{with} \qquad U(\boldsymbol{x}|\boldsymbol{x}^r, \mathcal{K}) = \sum_{t=1}^{T} -A_t^{\mathcal{K}}(\boldsymbol{x}^r) \log(A_t^{\mathcal{K}}(\boldsymbol{x})) \tag{3}$$

where $A_t^{\mathcal{K}}(\boldsymbol{x})$ denotes the area of the triangle t in a mesh with position $\boldsymbol{x}$, $\boldsymbol{x}^r$ is the most likely position of the mesh nodes (in the remainder called *reference* position), and the parameter β controls how far the mesh nodes can deviate from this reference position. Furthermore, having no specific prior knowledge about the values of the label probabilities $\boldsymbol{\alpha}$, we use a flat prior: $p(\boldsymbol{\alpha}) \propto 1$.

In a Bayesian setting, assessing the Maximum A Posteriori (MAP) parameters $\{\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{x}}^1, \ldots, \widehat{\boldsymbol{x}}^M\}$ involves minimizing

$$\sum_{m=1}^{M} -\log p(L_m|\boldsymbol{\alpha}, \boldsymbol{x}^m, \mathcal{K}) - \log p(\boldsymbol{x}^m|\beta, \boldsymbol{x}^r, \mathcal{K}). \tag{4}$$

We alternatively 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 atlas. Since the gradient of equation 4 with respect to $\boldsymbol{x}^m$ is given by analytical expressions, we perform this registration by gradient descent. 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. At each iteration, we calculate weights that associate each pixel in each example with each of the nodes attached to the triangle the pixel falls in

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

and update the parameters in each node n accordingly:

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

2.2 Second level of inference

The results of the atlas parameter estimation scheme described in section 2.1 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 $\widehat{\beta}$ by maximizing

$$p(L_1,\ldots,L_M|\beta,\boldsymbol{x}^r,\mathcal{K}) = \int_{\boldsymbol{\alpha}} \left(\prod_{m=1}^{M} p(L_m|\boldsymbol{\alpha},\beta,\boldsymbol{x}^r,\mathcal{K}) \right) p(\boldsymbol{\alpha}) \mathrm{d}\boldsymbol{\alpha} \qquad (5)$$

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

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. Using a similar Laplace approximation for the prior $p(\boldsymbol{x}^m|\beta,\boldsymbol{x}^r,\mathcal{K})$ in combination with the pseudo-likelihood approximation, and ignoring interdependencies between neighboring mesh nodes, we obtain[1]

$$p(L_m|\boldsymbol{\alpha},\beta,\boldsymbol{x}^r,\mathcal{K}) \;\simeq\; p(L_m|\boldsymbol{\alpha},\boldsymbol{x}_\alpha^m,\mathcal{K}) \cdot \prod_{n=1}^{N} O_n^m \qquad (6)$$

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

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

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

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 to 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 Newton's method to carry out the actual optimization.

Plugging equation 6 into equation 5, approximating $\boldsymbol{x}_\alpha^m$ and the factors O_n^m by their values at $\boldsymbol{\alpha} = \widehat{\boldsymbol{\alpha}}$, introducing the EM algorithm's weights, and using Stirling's approximation, $x! \simeq x^x e^{-x}$, we finally obtain

$$p(L_1,\ldots,L_M|\beta,\boldsymbol{x}^r,\mathcal{K}) \simeq \prod_{m=1}^{M}\prod_{n=1}^{N} \widehat{O}_n^m \cdot \prod_{n=1}^{N} \frac{(K{-}1)!\,\widehat{N}_n!}{(\widehat{N}_n{+}K{-}1)!} \cdot \prod_{m=1}^{M} p(L_m|\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{x}}^m,\mathcal{K}) \qquad (7)$$

[1] Here, D^2_θ denotes a matrix of second derivatives, or Hessian

where $\widehat{N}_n = \left[\sum_{m=1}^{M}\sum_{i=1}^{I}\widehat{W}_{i,n}^{m}\right]$ denotes the total number of pixels associated with node n at the MAP parameters $\{\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{x}}^1, \ldots, \widehat{\boldsymbol{x}}^M\}$. Equipped with equation 7, we assess the MAP estimate $\widehat{\beta}$ using a line search algorithm.

2.3 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 $p(L_1, \ldots, L_M|\boldsymbol{x}^r, \mathcal{K}) = \int_\beta p(L_1, \ldots, L_M|\beta, \boldsymbol{x}^r, \mathcal{K})\ p(\beta)\ \mathrm{d}\beta$, which can be approximated, using Laplace's method, by

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

The first factor is typically overwhelmed by the second one, so we will ignore it and compare alternative models based on equation 7, evaluated at the MAP estimate $\widehat{\beta}$.

While it is straightforward to compare models using equation 7, finding the mesh with connectivity and reference position that explicitly maximizes equation 7 is another matter. In this paper, we start from a dense regular triangular mesh, and use a mesh simplification technique borrowed from the computer graphics literature [10]. The technique yields increasingly coarse meshes by iteratively unifying two adjacent mesh nodes into a single node using a so-called edge collapse operation; each iteration removes the edge that yields the highest increase (or lowest decrease) in equation 7. For each edge collapse operation, we optimize the reference position $\boldsymbol{x}_n^r$ of the unified node n with respect to equation 7 using Powell's direction set. Finally, from the resulting hierarchy of meshes, we retain the one that yields the highest likelihood as evaluated by equation 7.

2.4 Description length interpretation

Given the central role of equation 7 in this paper, it is instructive to write it 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 $\{\mathcal{K}, \boldsymbol{x}^r\}$ is used. Following Shannon theory, this length is $-\log_2 p(L_1, \ldots, L_M|\boldsymbol{x}^r, \mathcal{K})$, which we approximate by (equation 7)

$$-\sum_{m=1}^{M}\sum_{n=1}^{N}\log_2 \widehat{O}_n^m \quad - \quad \sum_{n=1}^{N}\log_2 \frac{(K{-}1)!\ \widehat{N}_n!}{(\widehat{N}_n{+}K{-}1)!} \quad - \quad \sum_{m=1}^{M}\log_2 p(L_m|\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{x}}^m, \mathcal{K}).$$

According to the three terms, the message can be imagined as being subdivided into three blocks. Prior to starting the communication, the transmitter estimates

the MAP estimates $\{\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{x}}^1, \ldots, \widehat{\boldsymbol{x}}^M\}$ as laid out in section 2.1. It then sends a message block that encodes, for each label image, the position of each mesh node (first term). Subsequently, a message block is sent that encodes the prior probabilities in each mesh node (second term), after which the actual data can be encoded using the model with the MAP parameters (third term).

3 Experiments

We evaluated the performance of our mesh-based atlas models on 2-D training data derived from publicly available manual annotations[2]. In a first experiment, we derived probabilistic atlases from 3 training datasets with 4 labels for three different model sub-groups (figure 1). The first sub-group uses a regular triangular mesh model and prevents the mesh nodes from moving away from the reference position $\boldsymbol{x}^r$ by setting $\beta = 0$ throughout. Within this sub-group we measured the description length as the resolution of the mesh, i.e. the distance between the mesh nodes, is varied, and retained the best model (top right in figure 1). The second sub-group also uses a regular triangular mesh model, but explicitly searches for the best β as the resolution of the mesh is varied (bottom left of figure 1). Finally, we used the mesh simplification procedure outlined in section 2.3 (bottom right of figure 1). The first sub-group did not perform well compared to the other groups: while it saves bits by not needing to encode the mesh node positions, it does not model the data very well, resulting in a long data message block. Clearly, letting β vary, as in the second sub-group, pays off, but an even shorter description length is obtained using the mesh simplification procedure.

In a second experiment, we added 15 more training datasets and searched again for the best resolution within the first sub-group (top part of figure 2). Comparing this with the result based on only three training datasets reveals that the number of mesh nodes has increased. Note that using a higher mesh resolution is akin to reducing the amount of blurring in the resulting atlas; Bayesian inference thus automatically and quantitatively determines the "correct" amount of blurring that should be applied. Also note that, while the mesh does have more nodes when eighteen training datasets are used, there are still far fewer mesh nodes than there are pixels (around 25 times fewer). The bottom part of figure 2 shows the atlas obtained by pixel-wise averaging, i.e. by setting the mesh resolution so high that each node corresponds to exactly one pixel. From the message length representation, it is clear that this is not a good model: there are far too many model parameters, resulting in a severely overfitted model.

In a final experiment, we searched for the best model in the same three sub-groups used before, using 9 training datasets containing 11 labels (figure 3). Again, the best atlas is obtained by mesh simplification.

[2] The Internet Brain Segmentation Repository, http://www.cma.mgh.harvard.edu/ibsr

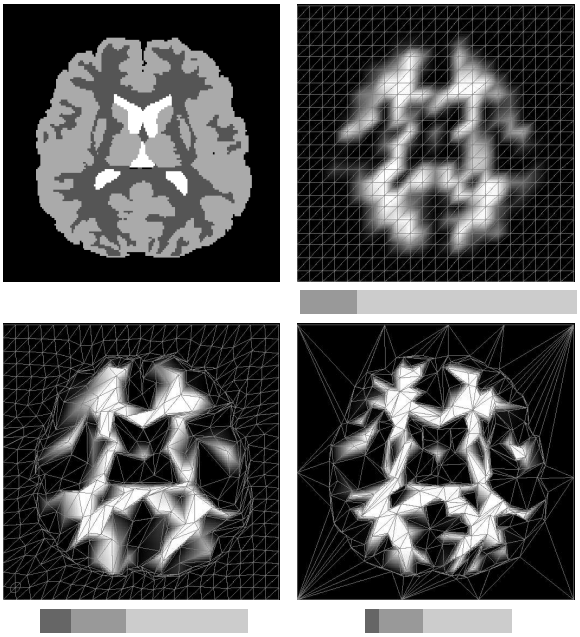

Fig. 1. Optimal probabilistic atlases constructed from three training datasets with four labels, for three subgroups of our mesh-based atlas models (see text for details). The top left figure depicts the first dataset of the training data; the two lower figures show atlases warped onto this dataset. Under each atlas is depicted a schematic view of the shortest message that encodes the training data: dark gray indicates the node position message block, intermediate gray represents the prior probabilities message block, and bright gray stands for the data message block.

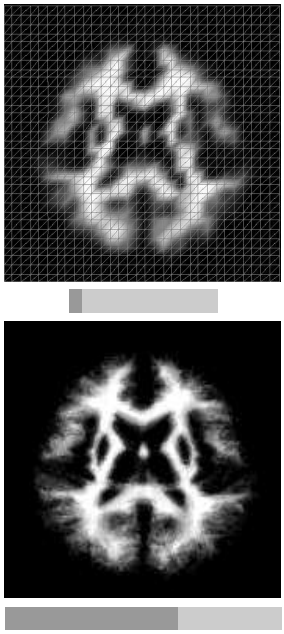

Fig. 2. "Average" atlas for eighteen training datasets at optimal resolution (top), and at pixel resolution (bottom).

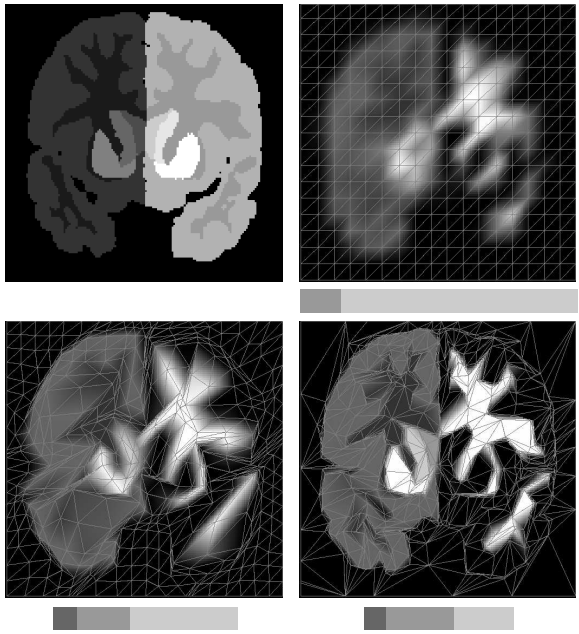

Fig. 3. Optimal probabilistic atlases based on nine images with eleven labels. The left side of the brain has been color-coded in the atlases for visualization purposes.

4 Discussion

While this paper concentrated on constructing prior probability distributions from training data, in practical segmentation scenarios the resulting atlases need to be aligned with the MRI data at hand before segmentation can commence. This requires that appropriate intensity distribution models are associated with each label. The gradient of the mesh node positions is then given in analytical form through equation 2, so that atlas-to-image registration is straight-forward to implement. Similar registration techniques that directly align priors with MR brain images have been described in [4, 7, 5].

Our future work will concentrate on implementing our atlas construction techniques in 3 dimensions, using tetrahedral rather than triangular meshes. We also plan to explore even more compact representations by explicitly encoding which subset of labels can occur at any given location throughout the brain.

References

1. K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens. Automated model-based tissue classification of MR images of the brain. *IEEE Transactions on Medical Imaging*, 18(10):897–908, October 1999.
2. K.M. Pohl, W.M. Wells, A. Guimond, K. Kasai, M.E. Shenton, R. Kikinis, W.E.L. Grimson, and S.K. Warfield. Incorporating non-rigid registration into Expectation Maximization algorithm to segment MR images. In *MICCAI 2002*, volume 2488 of *LNCS*, pages 564–571. Springer-Verlag, 2002.
3. A.P. Zijdenbos, R. Forghani, and A.C. Evans. Automatic "pipeline" analysis of 3-D MRI data for clinical trials: Application to multiple sclerosis. *IEEE Transactions on Medical Imaging*, 21(10):1280–1291, October 2002.
4. B. Fischl, D.H. Salat, A.J.W. van der Kouwe, N. Makris, F. Segonne, B.T. Quinn, and A.M. Dalea. Sequence-independent segmentation of magnetic resonance images. *NeuroImage*, 23:S69–S84, 2004.
5. J. Ashburner and K.J. Friston. Unified segmentation. *NeuroImage*, 26:839–851, 2005.
6. M. Prastawa, J.H. Gilmore, W. Lin, and G. Gerig. Automatic segmentation of MR images of the developing newborn brain. *Medical Image Analysis*, 9:457–466, 2005.
7. E. D'Agostino, F. Maes, D. Vandermeulen, and P. Suetens. Non-rigid atlas-to-image registration by minimization of class-conditional image entropy. In *MICCAI 2004*, volume 3216 of *LNCS*, pages 745–753. Springer-Verlag, 2004.
8. M. De Craene, A. du Bois d'Aische, B. Macq, and S.K. Warfield. Multi-subject registration for unbiased statistical atlas construction. In *MICCAI 2004*, volume 3216 of *LNCS*, pages 655–662. Springer-Verlag, 2004.
9. P. Lorenzen, M. Prastawa, B. Davis, G. Gerig, E. Bullitt, and S. Joshi. Multi-modal image set registration and atlas formation. *Medical Image Analysis*, 2006. In press.
10. H. Hoppe. Progressive meshes. In *ACM SIGGRAPH 1996*, pages 99–108, 1996.