Automated Brain Structure Segmentation Based on Atlas Registration and Appearance Models

Fedde van der Lijn*, Marleen de Bruijne, Stefan Klein, Tom den Heijer, Yoo Y. Hoogendam, Aad van der Lugt, Monique M. B. Breteler, and Wiro J. Niessen

***Abstract*—Accurate automated brain structure segmentation methods facilitate the analysis of large-scale neuroimaging studies. This work describes a novel method for brain structure segmentation in magnetic resonance images that combines information about a structure's location and appearance. The spatial model is implemented by registering multiple atlas images to the target image and creating a spatial probability map. The structure's appearance is modeled by a classifier based on Gaussian scale-space features. These components are combined with a regularization term in a Bayesian framework that is globally optimized using graph cuts. The incorporation of the appearance model enables the method to segment structures with complex intensity distributions and increases its robustness against errors in the spatial model. The method is tested in cross-validation experiments on two datasets acquired with different magnetic resonance sequences, in which the hippocampus and cerebellum were segmented by an expert. Furthermore, the method is compared to two other segmentation techniques that were applied to the same data. Results show that the atlas- and appearance-based method produces accurate results with mean Dice similarity indices of 0.95 for the cerebellum, and 0.87 for the hippocampus. This was comparable to or better than the other methods, whereas the proposed technique is more widely applicable and robust.**

***Index Terms*—Atlas registration, brain structures, graph cuts, MRI, pattern recognition, segmentation.**

I. Introduction

SEVERAL neurological disorders like Parkinson's disease, schizophrenia, and Alzheimer's disease are believed to cause changes in the shape and size of brain structures. Due to its excellent soft-tissue contrast, magnetic resonance imaging (MRI) has become an important tool for studying the etiology and consequences of these diseases [1]–[3]. Extracting size or shape information from MR images requires accurate and reliable delineation of brain structures. Although manual segmentation is still the gold standard for the analysis of image data, automated segmentation can help decrease the workload and increase reproducibility.

Many methods for automated brain structure segmentation have been introduced in the past decade. Among these, atlas-based segmentation techniques have arguably been the most successful. In this approach a manually labeled atlas image is registered to an unlabeled target image. The resulting deformation field is then used to warp the atlas labels to the target's coordinate system. Especially when registration errors are compensated for by adding a statistical intensity model, this approach can produce very accurate and robust results [4]–[9].

Many of these atlas- and intensity-based methods use a global intensity model to determine whether a voxel belongs to the foreground or background class [5]–[9]. However, most brain structures have one or more neighbors with similar intensity characteristics, which results in partially overlapping foreground and background distributions. Background voxels with foreground intensities (or vice versa) will, therefore, be mislabeled by the intensity model, unless the atlas information is very strong. As a result atlas- and intensity-based methods are unsuited to segment structures with complex, spatially varying intensity patterns like the cerebellum (Fig. 1). A large number of background voxels at the interface with the cerebrum and the brainstem will be considered as foreground. This limits the applicability of atlas- and intensity-based techniques. Structures with relatively simple intensity patterns like the hippocampus show limited overlap, and can, therefore, usually be accurately segmented. But when applied to the hippocampus, atlas- and intensity-based methods remain vulnerable to registration errors that push the atlas into background areas like the enthorinal cortex, amygdala, or parahippocampal gyrus (Fig. 1).

One way to overcome these limitations is by combining atlas registration with a local intensity model that can describe spatially varying intensity distributions [4], [10]. More recent work has shown the potential of modeling local image appearance of brain structures with high-dimensional feature vectors of Haar filters or Gaussian derivatives at different scales [11]–[16].

Manuscript received December 08, 2010; revised March 23, 2011; accepted August 18, 2011. Date of publication September 19, 2011; date of current version February 03, 2012. The work of F. van der Lijn was supported by the Netherlands Organization for Scientific Research (NWO). The work of M. de Bruijne was supported in part by NWO and in part by the Danish Council for Strategic Research (DSF). *Asterisk indicates corresponding author.*

*F. van der Lijn is with the Departments of Medical Informatics and Radiology, Erasmus MC, 3000 CA Rotterdam, The Netherlands (e-mail: f.vanderlijn@erasmusmc.nl).

M. de Bruijne is with the Departments of Medical Informatics and Radiology, Erasmus MC, 3000 CA Rotterdam, The Netherlands, and with the Department of Computer Science, University of Copenhagen, DK-2100 Copenhagen, Denmark (e-mail: marleen.debruijne@erasmusmc.nl).

S. Klein is with the Departments of Medical Informatics and Radiology, Erasmus MC, 3000 CA Rotterdam, The Netherlands (e-mail: s.klein@erasmusmc.nl).

T. den Heijer is with the Department of Neurology, Sint Franciscus Gasthuis, 3045 PM Rotterdam, The Netherlands, and with the Department of Epidemiology, Erasmus MC, 3000 CA Rotterdam, The Netherlands (e-mail: t.denheijer@sfg.nl).

Y. Y. Hoogendam is with the Department of Epidemiology, Erasmus MC, 3000 CA Rotterdam, The Netherlands (e-mail: y.hoogendam@erasmusmc.nl).

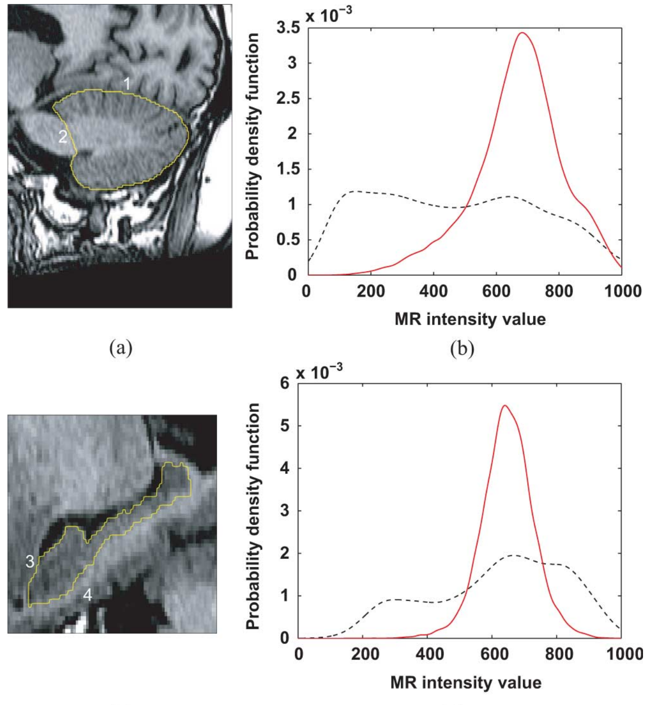
A. van der Lugt is with the Department of Radiology, Erasmus MC, 3000 CA Rotterdam, The Netherlands (e-mail: a.vanderlugt@erasmusmc.nl).

M. M. B. Breteler is with the German Center for Neurodegenerative Diseases, 53175 Bonn, Germany (e-mail: m.breteler@erasmusmc.nl).

W. J. Niessen is with the Departments of Medical Informatics and Radiology, Erasmus MC, 3000 CA Rotterdam, The Netherlands, and Imaging Science and Technology, Faculty of Applied Science, Delft University of Technology, 2600 AA Delft, The Netherlands (e-mail: w.niessen@erasmusmc.nl).

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

Digital Object Identifier 10.1109/TMI.2011.2168420



(a) (b)

(c) (d)

Fig. 1. Manually segmented T1-weighted images of a cerebellum and a hippocampus (both in sagittal view) together with the intensity distributions sampled from the structure (solid red curves) and background (dashed black curves). The cerebellum distribution overlaps with the intensities of the cerebrum (1) and brainstem (2). In the hippocampus image background voxels from the amygdala (3) and parahippocampal gyrus (4) have similar intensities as the foreground. (a) Cerebellum. (b) Intensity distribution cerebellum. (c) Hippocampus. (d) Intensity distribution hippocampus.

In this paper, we present a segmentation method that combines atlas-based segmentation with multi-feature appearance models. The proposed method uses multiple atlas registrations to construct a spatial probability map that models the location of the brain structure in an unlabeled MR image. The appearance of the structure is described by a voxel classifier based on Gaussian scale-space features. The smoothness of the result is controlled with a regularization term. The spatial, appearance, and regularization terms are then combined in a posterior probability function that can be globally maximized using graph cuts [17], [18].

This work is closest related to the hybrid discriminative/generative segmentation method presented in [14], which describes a combination of an appearance model and a statistical shape model. The appearance model is implemented with a boosting technique that uses approximately 5000 image and location features to describe local image information derived from a manually labeled training set. The shape model encourages segmentations that have a similar shape as the training images. In contrast, in our method this role is performed by the spatial probability map, which benefits from the accuracy and robustness of multiple atlas registrations.

The method is evaluated by segmenting the cerebellum and hippocampus in two MRI datasets that were acquired with different scanners and sequences. We determined its accuracy by comparing the results to manual segmentations. The performance of the method was also compared with that of two other techniques based on atlas registration [19] and atlas registration plus an intensity model [7].

II. Method

The segmentation of an unlabeled target image is equivalent to finding the label field $\mathbf{f}$ with the maximum posterior probability given the image information $\mathbf{i}$

$$\hat{\mathbf{f}} = \arg\max_{\mathbf{f}} p(\mathbf{f}|\mathbf{i}). \tag{1}$$

As we will consider binary segmentations, $\mathbf{f}$ is a vector containing a label $f_m \in \{0, 1\}$ for every voxel m in the set $\mathcal{M}$ of voxel locations in the image. Vector $\mathbf{i}$ consists of the intensity values i_m for all voxel locations $\mathcal{M}$.

Explicitly modeling the joint posterior probability $p(\mathbf{f}|\mathbf{i})$ would be feasible only for the smallest of images because of the exponential amount of possible label configurations. However, we can simplify the computation of $p(\mathbf{f}|\mathbf{i})$ by assuming that the label f_m conditioned on the image intensities $\mathbf{i}$ depends only on the labels of its neighbors $n \in \mathcal{N}_m$. This allows us to approximate (1) as a Discriminative Random Field (DRF) with one- and two-voxel clique potentials [20], [21]

$$p(\mathbf{f}|\mathbf{i}) \approx \frac{1}{Z} \exp\left(\sum_{m\in\mathcal{M}} \lambda_1 A(f_m, \mathbf{i}) + \sum_{m\in\mathcal{M}} \sum_{n\in\mathcal{N}_m} I(f_m, f_n, \mathbf{i}) \right) \tag{2}$$

in which we shall assume the terminology of [20], [21] and call $A(f_m, \mathbf{i})$ the association potential and $I(f_m, f_n, \mathbf{i})$ the interaction potential. λ_1 is a weight parameter that scales the association potential with respect to the interaction potential. Z is a constant that normalizes the summed posterior probabilities of all possible label configurations to one. As we are only interested in the label configuration $\hat{\mathbf{f}}$ that gives the maximum posterior probability, we can disregard this term in the optimization of (2). The association potential is proportional to the log probability that a single voxel assumes a foreground or background label, given all intensity values of the image. This term is based on statistical models of the brain structure's location and appearance. The interaction potential models the relation between two neighboring voxels, given $\mathbf{i}$. In this work, it is implemented as a regularizer that promotes piecewise smooth segmentations.

In Section II-A, the association potential is described, which consists of three components: an appearance model, a spatial model, and a global prior term, which will be detailed in separate subsections. Subsequently, the interaction potential is described. Finally, the section is completed with an explanation of the methods used to compute the maximum a posteriori solution and to select the model parameters.

A. *Association Potential*

In this work, the association potential has the following form:

$$A(f_m, \mathbf{i}) = \ln p_A(f_m|\mathbf{i}) \tag{3}$$

in which p_A is the association probability function that can be written as

$$p_A(f_m|\mathbf{i}) = \frac{p_{gp}(f_m) \cdot p_{\text{model}}(f_m|\mathbf{i})}{Z_A}. \tag{4}$$

In this equation $p_{\text{model}}(f_m|\mathbf{i})$ is a probability function that contains the spatial and appearance models. $p_{gp}(f_m)$ is a global prior term which affects the probability of foreground voxel labels in the entire image. Z_A is the constant that normalizes the summed foreground and background probabilities

$$Z_A = p_{gp}(f_m = 0) \cdot p_{\text{model}}(f_m = 0|\mathbf{i}) + p_{gp}(f_m = 1) \cdot p_{\text{model}}(f_m = 1|\mathbf{i}).$$

The model term $p_{\text{model}}(f_m|\mathbf{i})$ is defined as

$$p_{\text{model}}(f_m|\mathbf{i}) = \frac{p_{\text{app}}(f_m|\boldsymbol{\xi}_m(\mathbf{i})) \cdot p_s^{\lambda_2}(f_m)}{Z_{\text{model}}} \tag{5}$$

in which $p_{\text{app}}(f_m|\boldsymbol{\xi}_m(\mathbf{i}))$ is an appearance model probability map, based on an F-dimensional vector $\boldsymbol{\xi}_m(\mathbf{i})$ of appearance features extracted from the image in the neighborhood of m. It describes the probability of label f_m, based on the appearance at location m. $p_s(f_m)$ is a spatial probability map, which records the probability of encountering label f_m at voxel location m according to a spatial model. λ_2 determines the balance between the spatial and appearance models. Z_{model} is the constant that normalizes the summed probabilities to one

$$Z_{\text{model}} = p_{\text{app}}(f_m = 0|\boldsymbol{\xi}_m(\mathbf{i})) \cdot p_s^{\lambda_2}(f_m = 0) + p_{\text{app}}(f_m = 1|\boldsymbol{\xi}_m(\mathbf{i})) \cdot p_s^{\lambda_2}(f_m = 1).$$

B. Appearance Model

The appearance model $p_{\text{app}}(f_m|\boldsymbol{\xi}_m(\mathbf{i}))$ is constructed for the unlabeled target image u by applying a k-nearest-neighbor (knn) voxel classifier operating in the F-dimensional feature space. The classifier is trained by extracting foreground and background samples from a set of J manually labeled training images $\mathcal{T} = \{t_1, \ldots, t_J\}$. The probability that a voxel with a feature vector $\boldsymbol{\xi}_m$ has class label f_m is given by

$$p_{\text{app}}^{(u;\mathcal{T})}(f_m|\boldsymbol{\xi}_m(\mathbf{i})) = \frac{k_{f_m}(\boldsymbol{\xi}_m(\mathbf{i})) + 1}{k + 2} \tag{6}$$

in which $k_{f_m}(\boldsymbol{\xi}_m(\mathbf{i}))$ counts the number of training samples with label f_m among the k nearest neighbors of the point $\boldsymbol{\xi}_m(\mathbf{i})$ in the F-dimensional feature space. The superscript $(u;\mathcal{T})$ is used throughout this paper whenever it is important to explicitly specify the target image and the training set. Since the knn classifier makes no assumptions about the distribution of $\boldsymbol{\xi}_m(\mathbf{i})$, it can model complex decision boundaries and has been shown to be effective in brain structure segmentation [11].

We used a moderated knn instead of the standard expression $k_{f_m}(\boldsymbol{\xi}_m(\mathbf{i}))/k$ to ensure that no voxel labels could be "vetoed" by the appearance model [22]. The foreground samples were obtained by random sampling from all voxels that were labeled as part of the structure of interest. An equal number of background samples was randomly extracted from a band around the manual segmentation. In all experiments k was set to 10, and we used a fast implementation based on approximate nearest neighbor searching (with an error bound ϵ of one) [23].

The appearance was modeled with Gaussian scale-space features. These were a Gaussian-filtered version of the original image, first- and second-order Gaussian derivatives in all axis directions, gradient magnitude, Laplacian, Gaussian curvature, and the three eigenvalues of the Hessian. Including the original intensity values, we used $1 + 16n_\sigma$ features for classification (with n_σ the number of scales). The features were independently standardized to zero mean and unit variance. Unlike other segmentation methods that use high-dimensional feature vectors (for example, [11] and [13]), we did not include location features; this information is introduced in the model by the spatial component.

A subset of the most relevant image descriptors was found with sequential forward feature selection, followed by sequential backward selection. In this process, features were added until the area under the ROC curve no longer increased. Then features were iteratively removed until the performance started to deteriorate. The feature selection was trained on one half of the J training images and its performance was estimated on the other half.

The foreground sample fraction, extent of the background sample area, and the feature scales, were chosen differently for the different image sets and brain structures for which the appearance model was constructed. Their values can be found in Section III-B.

C. Spatial Model

The spatial model was constructed by non-rigidly registering the J training images to the unlabeled target image u. The labels of the training images were then deformed and averaged to create a probability map $p_s(f_m)$

$$p_s^{(u;\mathcal{T})}(f_m) = \frac{1}{J} \sum_{t_j \in \mathcal{T}} a_m^{(u;t_j)}. \tag{7}$$

In this equation $a_m^{(u;t_j)} \in \{0, 1\}$ represents the atlas label of training image t_j deformed to the coordinate frame of target image u and interpolated at voxel location m.

All registrations were computed by first finding an affine transformation, followed by a nonrigid transformation parameterized by B-splines. The nonrigid registration step was computed in a multi-resolution fashion with increasing B-spline control point resolution. In all cases mutual information was used as similarity measure. The brain-structure-specific registration settings are discussed in Section III-C. All registrations were computed using the Elastix software [24].

D. Global Prior

As the appearance model is trained with an equal number of samples for foreground and background, the resulting classifier might not accurately reflect the prior class probabilities. Furthermore, the spatial model might exhibit under- or over-segmentation. The global prior term can compensate for these types of errors by increasing or decreasing the posterior probability of a foreground in the entire image

$$p_{gp}(f_m) = \begin{cases} \alpha, & \text{if } f_m = 0 \\ 1 - \alpha, & \text{if } f_m = 1 \end{cases} \tag{8}$$

with parameter α between 0 and 1. If α has a value of 0.5, the association probability p_A will be unaffected. An α larger than 0.5 decreases the probability of a foreground label for all voxel locations and consequently decreases the volume of the segmentation, whereas a smaller value increases the foreground probability and the resulting volume.

E. Interaction Potential

As the association potential for each voxel is independent of the others, the resulting segmentation could be noisy. To increase the probability of smoother segmentations the interaction potential was implemented as follows:

$$I(f_m, f_n, \mathbf{i}) = \begin{cases} 0, & \text{if } f_m = f_n \\ -\frac{1}{2} w_{m,n} B(\Delta\xi_{m,n}(\mathbf{i})), & \text{if } f_m \neq f_n. \end{cases} \quad (9)$$

In this expression $w_{m,n}$ is a distance weight equal to $d_{m,n} / \sum_{l \in \mathcal{N}_m} d_{m,l}$, with $d_{m,n}$ the distance between voxel locations m and n. $B(\Delta\xi_{m,n}(\mathbf{i}))$ is the penalty term for assigning different labels to voxels m and n, which is small if the appearance difference between the voxels is high. The appearance difference is modeled by the Euclidean distance $\Delta\xi_{m,n}(\mathbf{i})$ between the F-dimensional feature vectors $\boldsymbol{\xi}_m(\mathbf{i})$ and $\boldsymbol{\xi}_n(\mathbf{i})$. The penalty is given by a logistic function

$$B(\Delta\xi_{m,n}(\mathbf{i})) = \frac{1}{1 + \exp(\beta_0 + \beta_1 \Delta\xi_{m,n}(\mathbf{i}))} \quad (10)$$

in which β_0 and β_1 control the offset and slope of the term.

Equation (9) promotes smoother segmentations by decreasing the posterior probability of a label configuration in which neighboring voxels have different labels. However, if the feature space distance between neighboring voxels is large, we assume that they belong to different structures. In that case, the logistic function limits the reduction of the posterior probability. This model is a multi-feature version of the gradient-modulated Ising model commonly used in graph cut segmentation methods [7], [25].

F. Optimization and Parameter Learning

The posterior probability function $p(\mathbf{f}|\mathbf{i})$ described above is completely defined by the voxel classifier result p_{app} that models the appearance, the atlas registration result p_s that models the spatial probability, and a parameter vector $\boldsymbol{\theta}$. The latter holds the model's five free parameters: the association weight λ_1, the spatial model weight λ_2, the foreground threshold α, and the parameters of the logistic interaction model β_0 and β_1.

The optimal value of these parameters $\hat{\boldsymbol{\theta}}$ depends heavily on the quality of the appearance and spatial models, which is not known. However, the model quality can be estimated with cross-validation experiments using the manual segmentations of the training images. In this work specifically, we chose $\hat{\boldsymbol{\theta}}$ from a predefined range of values $\boldsymbol{\Theta}$ using exhaustive search, which is explained in more detail in Section III-B.

Once the parameters have been chosen, a maximum a posteriori (MAP) solution $\hat{\mathbf{f}}$ can be found by converting (1) to an equivalent energy function by taking the negative logarithm. As shown in [26] this function can be globally minimized using graph cuts [18]. In this work, we used the Maxflow algorithm to perform this optimization [27].

III. Experiments and Results

The method was tested by segmenting the cerebellum and the hippocampus in T1-weighted images. The cerebellum exhibits a complex, spatially varying intensity distribution, whereas the hippocampus has a simpler uniform intensity pattern. To assess the method's ability to handle different MR sequences we also segmented the hippocampus in Half-Fourier Acquisition Single-Shot Turbo Spin Echo (HASTE) images with a lower resolution. The segmentation accuracy was determined by computing overlap and distance measures with respect to manual segmentations in a leave-one-out experiment. The atlas- and appearance-based method was also compared to two alternative techniques. The subjects and image data are described in more detail in Section III-A. The parameter learning procedure and data-specific implementation details are given in Section III-B. The experiments are detailed in Section III-C. Finally, the results are described in Section III-D.

A. Image Data

We used two MR datasets from the Rotterdam Scan Study, an image-based longitudinal cohort study on diseases among the elderly. The subjects were taken from two different study cohorts and selected to cover the cohorts' variability in age, sex, and global brain size (as measured with an automated brain tissue segmentation method).

Set I consisted of 10 women and 8 men with a mean age of 74.2 ± 7.9 years. These images were made with a 1.5T General Electric scanner. We used a 3D T1-weighted sequence (inversion time 400 ms, repetition time 14.8 ms, time to echo 2.8 ms, 96 axial slices of 1.6 mm interpolated to 192 slices of 0.8 mm, acquisition matrix 416 × 256, field-of-view 250×250 mm). The final voxel size was 0.49 × 0.49 × 0.8 mm. The hippocampi and cerebellum in set I were segmented by one observer (Y. Y. Hoogendam) under supervision of a neurologist (T. den Heijer) and a neuro-radiologist (A. van der Lugt). These structures were delineated every other slice (in sagittal view for the cerebellum and in coronal view for the hippocampus). Linear interpolation was used to obtain segmentations for the skipped slices. We shall refer to the hippocampus segmentations of this set as I-HC and to the cerebellum segmentations as I-CRBL.

The low-resolution set II consisted of 9 women and 11 men with a mean age of 74.6±8.2 years. These images were acquired on a 1.5T Siemens scanner with a custom-made 3D HASTE sequence (inversion time 440 ms, repetition time 2800 ms, 128 contiguous sagittal slices of 1.25 mm, acquisition matrix 192 × 256, field-of-view 256 × 256 mm). Two HASTE modules were sequentially acquired after the inversion pulse (effective echo time of 29 ms and 440 ms) of which the first was used in this work. The final voxel size of these images was 1.25 × 1.0 × 1.0 mm. In these images the hippocampi were delineated on coronal slices by two raters. Fifteen images were segmented by an expert neurologist (T. den Heijer) and five by a trained observer (Y. Y. Hoogendam) under supervision of a neurologist (T. den Heijer). We shall refer to these images and their labels as the II-HC set. The images from both sets were corrected for nonuniformities using N3 [28].

B. Segmentation Procedure

The atlas- and appearance-based method was applied to the I-CRBL, I-HC, and the II-HC sets. These segmentations were performed in a leave-one-out experiment consisting of three steps. First for every image $t_j \in \mathcal{T}$ an appearance model $p_{\text{app}}^{(t_j;\mathcal{T}_j)}$

```
for $t_j \in \mathcal{T}$ do
    Construct appearance model $p_{app}^{(t_j;\mathcal{T}_j)}$
    Construct spatial model $p_s^{(t_j;\mathcal{T}_j)}$
end for
for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ do
    for $t_j \in \mathcal{T}$ do
        Compute posterior probability function
            $p^{(t_j;\mathcal{T}_j)}(\boldsymbol{\theta})$ based on $p_{app}^{(t_j;\mathcal{T}_j)}$, $p_s^{(t_j;\mathcal{T}_j)}$, and $\boldsymbol{\theta}$
        Compute MAP label configuration $\hat{\mathbf{f}}^{(t_j)}(\boldsymbol{\theta})$
        Compute $DSI(\hat{\mathbf{f}}^{(t_j)}(\boldsymbol{\theta}), \mathbf{g}^{(t_j)})$
    end for
end for
for $t_j \in \mathcal{T}$ do
    Compute
        $\overline{DSI}^{(t_j)}(\boldsymbol{\theta}) = 1/N \sum_{t_k \in \mathcal{T}_j} DSI(\hat{\mathbf{f}}^{(t_k)}(\boldsymbol{\theta}), \mathbf{g}^{(t_k)})$
    Find $\hat{\boldsymbol{\theta}}^{(t_j)} = \arg\max_{\boldsymbol{\theta}} \overline{DSI}^{(t_j)}(\boldsymbol{\theta})$
    Compute posterior function $p^{(t_j;\mathcal{T}_j)}(\hat{\boldsymbol{\theta}}^{(t_j)})$ based on
        $p_{app}^{(t_j;\mathcal{T}_j)}$, $p_s^{(t_j;\mathcal{T}_j)}$, and $\hat{\boldsymbol{\theta}}^{(t_j)}$
    Compute MAP label configuration $\hat{\mathbf{f}}^{(t_j)}(\hat{\boldsymbol{\theta}}^{(t_j)})$
end for
```

Fig. 2. Parameter learning procedure. See text for more details.

and a spatial model $p_s^{(t_j;\mathcal{T}_j)}$ was created using the remaining subjects' scans $\mathcal{T}_j = \mathcal{T} \setminus \{t_j\}$ as training images. Secondly, based on these models, segmentations $\hat{\mathbf{f}}^{(t_j)}(\boldsymbol{\theta})$ were computed for all parameter values $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and all target images $t_j \in \mathcal{T}$. We then measured the Dice similarity indices $DSI(\hat{\mathbf{f}}^{(t_j)}(\boldsymbol{\theta}), \mathbf{g}^{(t_j)})$ between $\hat{\mathbf{f}}^{(t_j)}(\boldsymbol{\theta})$ and the manual segmentations $\mathbf{g}^{(t_j)}$. This records the segmentation accuracy as a function of the parameters $\boldsymbol{\theta}$ and target image t_j. In the third step, the optimal parameters $\hat{\boldsymbol{\theta}}^{(t_j)}$ were selected for target t_j by finding the parameters that gave the highest mean similarity index computed over all other images $\mathcal{T}_j$. With these parameters the segmentation $\hat{\mathbf{f}}^{(t_j)}(\hat{\boldsymbol{\theta}}^{(t_j)})$ was computed. In this way, parameter learning for the segmentation of t_j was never based on spatial or appearance models constructed in the coordinate system of t_j. The whole procedure is summarized in Fig. 2.

To create the appearance model for I-CRBL, 1% of the manually labeled foreground voxels in the training images were sampled. The background samples were taken from a band up to 10 mm around the foreground. We used $n_\sigma = 4$ with equal logarithmic intervals between 0.5 and 5 mm. The I-HC appearance model was based on 5% of the foreground voxels and a background band of 4 mm. Five scales were used between 0.5 and 5 mm. The sampling parameters of II-HC were identical to those of I-HC, but because of the lower resolution of the images we used three scales between 1 and 5 mm.

The spatial models for I-CRBL and II-HC were based on mutual information computed over the entire image. For I-HC the registration was initialized with the deformation field computed for the I-CRBL set, and further refined in a region of interest around the hippocampus. The registration settings can be found in the Elastix parameter database.[1]

The interaction potential was based on a 26-voxel 3D neighborhood for the hippocampi. To reduce computation cost and memory requirement of the graph cut we used a 6-voxel 3D neighborhood for the cerebellum. For the same reason, separate subimages were created around the left and right hippocampus and the cerebellum after construction of the spatial and appearance models. These cropped images were based on bounding boxes around the thresholded spatial probability maps of the structures.

Computation time of a registration of one atlas image to the target image was approximately 10 CPU min on a node of a 64-bit Linux cluster. As a result the construction of the spatial model took 3 CPU h for I-CRBL, 6 CPU h for I-HC, and 3.5 CPU h for II-HC. The atlas registrations were performed in parallel to reduce computation time. The computation of the appearance model took approximately 0.5 CPU h per image of I-HC, 1 CPU h per image of I-CRBL, and 2 CPU min per image of II-HC. Constructing and maximizing the posterior probability function was done within a second for the hippocampi and in 2 min for the cerebellum on a desktop computer. The parameter learning took about four days for I-CRBL, two days for I-HC, and one day for II-HC.

Separating the I-CRBL images in background, left and right side voxels is a three-class segmentation problem that can not be globally solved with graph cuts optimization [18]. Therefore, we first applied the proposed method to label all voxels as cerebellum and background. All cerebellum voxels were then classified as left and right using the atlas. All DSI scores obtained during parameter learning were computed for the entire cerebellum.

C. Experiments

The atlas- and appearance-based method was validated by comparing the results of the leave-one-out segmentations to the manual labelings. We used the following volumetric quality measures. The Dice similarity index (DSI), defined as

$$\mathrm{DSI} = \frac{2V(\mathbf{f} \cap \mathbf{g})}{V(\mathbf{f}) + V(\mathbf{g})} \tag{11}$$

the Jacquard similarity index (JSI), given by

$$\mathrm{JSI} = \frac{V(\mathbf{f} \cap \mathbf{g})}{V(\mathbf{f} \cup \mathbf{g})}, \tag{12}$$

the relative volume difference (RV), defined by

$$\mathrm{RV} = \frac{V(\mathbf{f}) - V(\mathbf{g})}{V(\mathbf{g})} \tag{13}$$

and the volumetric, two-way random, absolute agreement, intraclass correlation coefficient (ICC) [29] between $V(\mathbf{g})$ and $V(\mathbf{f})$. In these expressions, $V(\mathbf{f})$ and $V(\mathbf{g})$ are the volumes of the automated segmentation $\mathbf{f}$ and the manual segmentation $\mathbf{g}$.

We also computed two surface-based measures: the maximum and mean surface distance $D_{\max}$ and D_{mean}. The maximum distance is given by

$$D_{\max} = \max\{\delta(\mathbf{f}, \mathbf{g}), \delta(\mathbf{g}, \mathbf{f})\} \tag{14}$$

with $\delta(\mathbf{f}, \mathbf{g})$ a set that contains the distances between every surface voxel in automated segmentation $\mathbf{f}$, and the closest surface voxel in the manual segmentation $\mathbf{g}$. The mean surface distance is defined by

$$D_{\mathrm{mean}} = \frac{\bar{\delta}(\mathbf{f}, \mathbf{g}) + \bar{\delta}(\mathbf{g}, \mathbf{f})}{2} \tag{15}$$

[1] http://elastix.isi.uu.nl/wiki

TABLE I
EVALUATION MEASURES FOR THE ATLAS- AND APPEARANCE-, ATLAS- AND INTENSITY-, AND ATLAS-BASED METHODS. FOR ALL SCORES EXCEPT THE ICC MEAN, STANDARD DEVIATION, AND RANGE ARE LISTED. P-VALUES WERE COMPUTED USING A KRUSKAL-WALLIS SIGNED RANK TEST WHICH TESTS THE HYPOTHESIS THAT ALL THREE MEDIAN SCORES ARE EQUAL

	atlas&appearance	**atlas&intensity**	**atlas**	**KW-test**
I-CRBL				
DSI	0.954±0.008 [0.925 ; 0.963]		0.937±0.013 [0.900 ; 0.953]	$p < 0.001$
JSI	0.911±0.015 [0.861 ; 0.928]		0.882±0.023 [0.818 ; 0.911]	$p < 0.001$
RV	0.003±0.039 [-0.119 ; 0.099]		0.007±0.057 [-0.145 ; 0.121]	p = 0.55
ICC [95%CI]	0.912 [0.853 ; 0.954]		0.817 [0.671 ; 0.902]	-
D_{mean}, mm	0.50±0.10 [0.38 ; 0.79]		0.68±0.15 [0.50 ; 1.09]	$p < 0.001$
D_{max}, mm	7.35±3.10 [3.28 ; 16.91]		7.35±1.92 [3.90 ; 10.74]	p = 0.68
I-HC				
DSI	0.870±0.017 [0.829 ; 0.899]	0.867±0.018 [0.814 ; 0.907]	0.858±0.017 [0.817 ; 0.892]	p = 0.008
JSI	0.771±0.026 [0.708 ; 0.816]	0.766±0.028 [0.686 ; 0.830]	0.752±0.027 [0.691 ; 0.804]	p = 0.008
RV	0.031±0.092 [-0.122 ; 0.244]	0.016±0.096 [-0.133 ; 0.273]	0.000±0.079 [-0.156 ; 0.194]	p = 0.34
ICC [95%CI]	0.633 [0.391 ; 0.793]	0.609 [0.354 ; 0.779]	0.724 [0.522 ; 0.849]	-
D_{mean}, mm	0.34±0.06 [0.23 ; 0.53]	0.35±0.64 [0.21 ; 0.58]	0.36±0.06 [0.24 ; 0.52]	p = 0.35
D_{max}, mm	3.69±0.99 [1.93 ; 5.45]	3.86±0.97 [2.33 ; 6.33]	3.53±0.93 [2.18 ; 5.52]	p = 0.34
II-HC				
DSI	0.865±0.022 [0.818 ; 0.908]	0.864±0.028 [0.786 ; 0.910]	0.835±0.035 [0.736 ; 0.892]	$p < 0.001$
JSI	0.762±0.034 [0.692 ; 0.831]	0.761±0.043 [0.647 ; 0.834]	0.718±0.051 [0.582 ; 0.805]	$p < 0.001$
RV	0.011±0.109 [-0.218 ; 0.268]	0.015±0.116 [-0.226 ; 0.255]	0.028±0.167 [-0.209 ; 0.412]	p = 0.54
ICC [95%CI]	0.797 [0.647 ; 0.887]	0.733 [0.548 ; 0.850]	0.485 [0.205 ; 0.691]	-
D_{mean}, mm	0.38±0.08 [0.27 ; 0.62]	0.38±0.09 [0.25 ; 0.69]	0.46±0.11 [0.27 ; 0.77]	$p < 0.001$
D_{max}, mm	4.89±1.77 [2.56 ; 9.39]	5.02±1.63 [2.36 ; 9.01]	4.80±1.61 [2.50 ; 8.95]	p = 0.79

with $\bar{\delta}(\mathbf{f},\mathbf{g})$ the mean of set $\delta(\mathbf{f},\mathbf{g})$ computed over all surface voxels of $\mathbf{f}$.

To ascertain whether the multi-feature appearance model improves results compared to a model based on MR intensities only, we also segmented the I-HC and II-HC images with the atlas- and intensity-based method published in [7]. This method combines a spatial model, an MR intensity model, and a regularizer in an energy framework that is optimized by graph cuts. I-CRBL was not segmented with this method because its intensity model cannot adequately separate the structure's foreground and background intensities. The results were compared with the manual segmentations using the quality measures listed above.

The atlas- and intensity-based segmentations were obtained using the same spatial model as the atlas and appearance results. The intensity model for the target image t_j was based on a Parzen classifier trained on intensity values extracted from the remaining manually labelled images $\mathcal{T}_j$. Finally, we used an identical regularizer as in [7]. The model described in [7] did not include a global prior p_{gp}, so we added a similar term to the atlas- and intensity-based method. The resulting model had three free parameters (equivalent to the λ_1, λ_2, and α), which were optimized in the way described in Section III-B.

To assess the added value of the appearance model and the interaction potential, we also compared the performance of the proposed method to that of a multi-atlas-based segmentation [19]. This method was applied to all three datasets and validated using the same quality measures.

The atlas-based segmentations were computed by thresholding the spatial model $p_s^{(t_j;\mathcal{T}_j)}$ at value α. This threshold value was chosen based on the training data using a similar procedure as described in Section III-B. We chose to select the threshold based on the training data instead of using a fixed value of 0.5, to make the results better comparable to the proposed method with its global prior p_{gp} term.

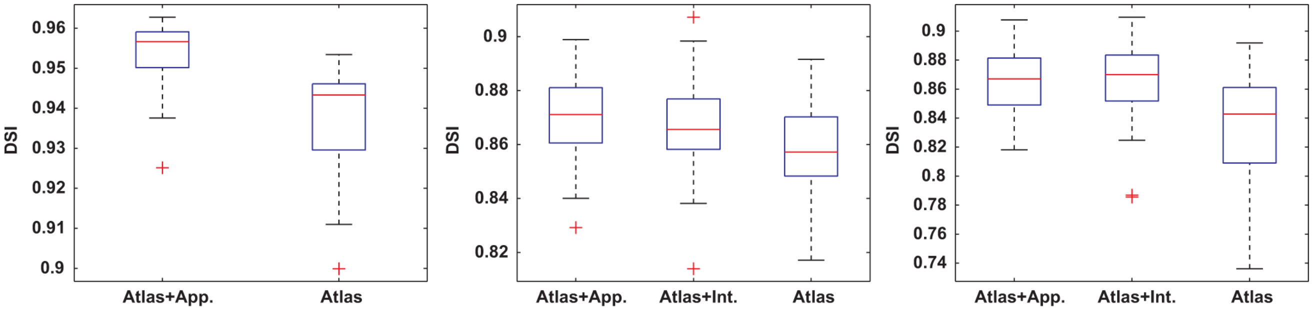
All scores are reported as mean ± standard deviation $[\min; \max]$. These statistics were computed over the left- and right-side structures of all images, so N was 36 for the I-HC and I-CRBL sets, and 40 for the II-HC set. We used Kruskal–Wallis signed rank tests to ascertain whether the atlas- and appearance-, atlas- and intensity-, and atlas-based methods had equal median scores. Additionally, the volume estimates were evaluated by plotting the automatically measured volumes against the manual volumes, and fitting a linear model through this data using linear regression.

Furthermore, we also computed the method's accuracy when using a smaller training set. For this purpose, set I was randomly divided in two folds of nine subjects each. One fold was used as training set to segment the other fold and evaluation scores were computed. Subsequently, training and test sets were switched to obtain quality measures for the other nine subjects. These cross-validation experiments were then repeated for four alternative combinations of two folds and averaged over the five draws. The resulting summary scores were averaged over the 36 left- and right-side cerebella and hippocampi, and compared to the results of the previous experiment using a Wilcoxon signed rank test. This should give an indication whether the proposed method can also produce accurate results with a smaller training set.

Finally, to assess the robustness of the parameter learning procedure, we analyzed how the accuracy of the segmentations depended on the parameters. We visualized this relation making plots of the average Dice similarity index as a function of two parameters, while keeping the other three parameters at their optimal values. These optimal settings were defined as the parameter values that gave the highest average DSI. This experiment was performed for the I-CRBL and I-HC sets, which resulted in two sets of plots. As a reference we computed the maximum possible DSI by overtraining. This was done by selecting for each target image the parameter settings that gave the highest DSI.

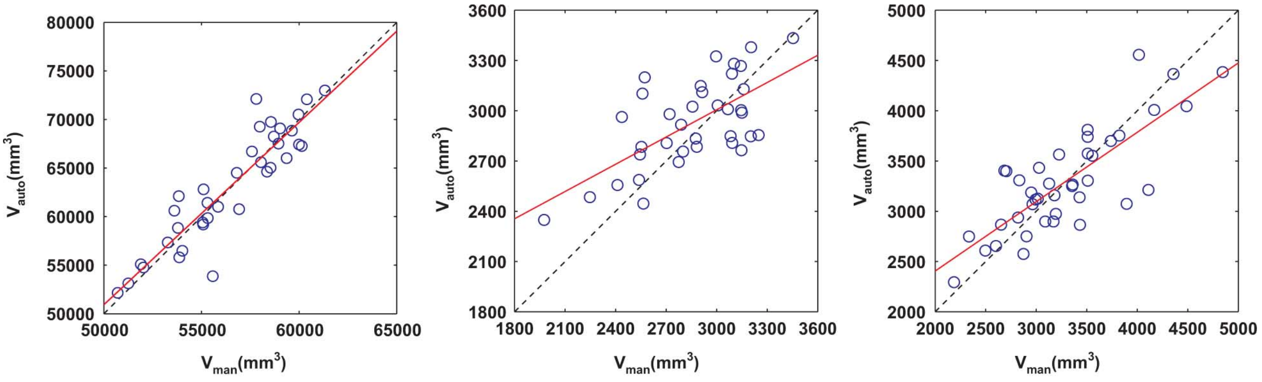
D. Results

Table I shows the quality scores of the atlas- and appearance-, atlas- and intensity-, and the atlas-based models for all three val-



(a) (b) (c)

Fig. 3. Boxplots of the different methods' DSI scores measured in the three validation sets. (a) I-CRBL. (b) I-HC. (c) II-HC.



(a) (b) (c)

Fig. 4. Scatterplots of the volumes measured with the atlas- and appearance-based method versus the manually measured volumes for the three validation sets. The regression lines are shown in solid red and the perfect segmentation as a dotted black line. (a) I-CRBL. (b) I-HC. (c) II-HC.

idation sets. Fig. 3 compares the different methods' DSI scores per dataset. When looking at the I-CRBL, the atlas- and appearance-based method yielded higher scores across the board compared to using the atlas alone. The performance of all three methods is very high on the I-HC set, although the atlas-based method never produces the best result. In the II-HC set the atlas- and intensity- and atlas- and appearance-based methods both perform better than the atlas-based method, but the difference between these two methods is small.

Fig. 4 shows scatter plots of the volume of the manual segmentation versus that of the atlas- and appearance-based segmentation. The regression coefficients of the linear model fitted on the measurements by all methods are shown in Table II. The atlas- and appearance-based model generally has the slopes that are closest to one, but all the automated methods have the tendency to underestimate large, and overestimate small volumes for the I-HC and II-HC datasets.

Fig. 5 shows the probability of encountering a foreground voxel in the two images from Fig. 1 according to the appearance model. Compared to the one-dimensional intensity distributions of Fig. 1, the foreground and background appearance distributions exhibit a much smaller overlap. The appearance model recognizes neighboring structures like the brainstem and the amygdala as background in Fig. 5(b) and (d). The majority of false positives are found further away from the structures, outside the band from which the background appearance was sampled. But these errors are compensated for by the spatial model, which will assign a zero foreground probability to these voxels.

TABLE II
Coefficients of the Linear Models That Map the Manual Volume to Volume Measured With the Atlas- and appearance-, Atlas- and intensity-, and Atlas-Based Methods

	atlas&appearance	atlas&intensity	atlas
I-CRBL			
Intercept, ml	4.0		9.0
(CI%95)	(-5.4;13.4)		(-4.4;22.5)
Slope	0.94		0.86
(CI%95)	(0.79;1.09)		(0.65;1.08)
I-HC			
Intercept, ml	1.4	1.4	1.1
(CI%95)	(0.8;2.0)	(0.8;2.1)	(0.6;1.7)
Slope	0.54	0.51	0.60
(CI%95)	(0.32;0.76)	(0.28;0.74)	(0.41;0.79)
II-HC			
Intercept, ml	1.0	1.5	2.2
(CI%95)	(0.5;1.6)	(1.0;2.0)	(1.7;2.8)
Slope	0.69	0.51	0.32
(CI%95)	(0.52;0.86)	(0.40;0.69)	(0.16;0.49)

Visual comparison of the results showed that the addition of an intensity- or appearance-based component corrects small under- and oversegmentations caused by registration errors (see Fig. 6). But the atlas- and intensity-based method has difficulties dealing with cases where the registration crosses over to neighboring gray matter regions. As can be seen in Fig. 7(c), the intensity model mistakes these areas for parts of the foreground, which further deteriorates the results. The appearance model recognized that these areas are background and corrected most of these errors [Fig. 7(b)] This accounts for the removal of the two outliers shown in Fig. 3(c).

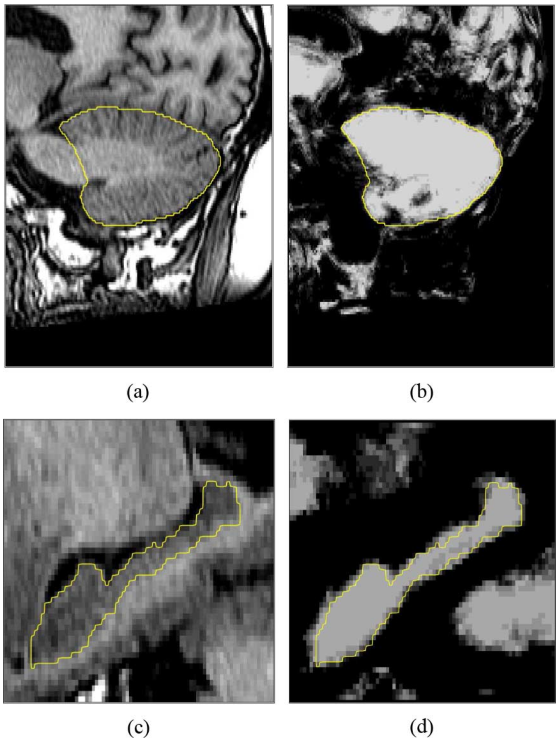


Fig. 5. Manually segmented T1-weighted images of a cerebellum and a hippocampus (sagittal view) together with the foreground probability according to the appearance model (higher intensities represent higher probabilities). (a) Cerebellum. (b) Appearance model cerebellum. (c) Hippocampus. (d) Appearance model hippocampus.

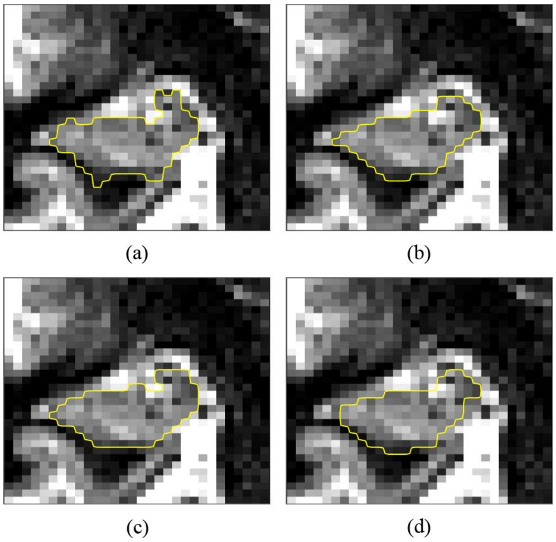


Fig. 6. Coronal slice through the segmentations of the same subject from the II-HC set. (a) Manual. (b) Atlas and appearance. (c) Atlas and intensity. (d) Atlas.

However, in the absence of large registration errors the effect of the intensity and appearance components are comparable. This is especially apparent in set II-HC, where in a small majority of cases the atlas and intensity model outperforms the atlas and appearance model [Fig. 3(c)]. The significant p-value of the Kruskal–Wallis test of the DSI scores listed for this validation set is purely due to the differences between the atlas-based method and the other methods. Unsurprisingly, a post-hoc Wilcoxon signed rank test shows no significant difference between the DSI scores of the atlas- and intensity-, and atlas- and appearance-based methods.

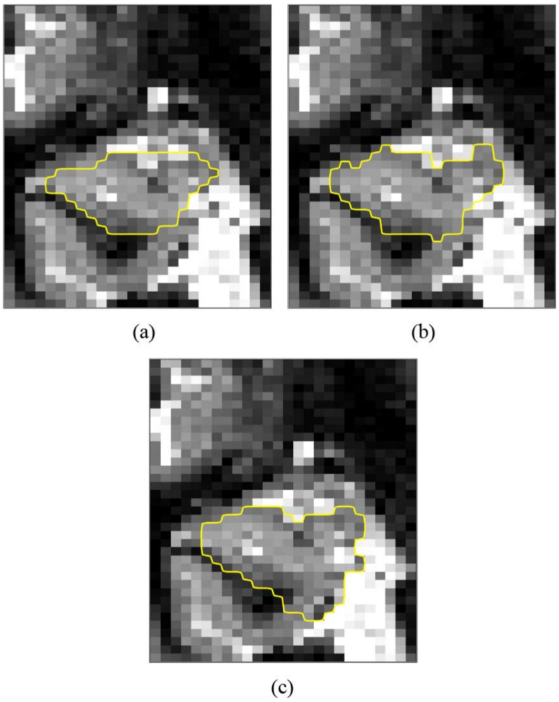


Fig. 7. Coronal slice taken from the II-HC set showing a large over–segmentation by the atlas- and intensity-based model. The atlas- and appearance-based model avoids these errors. (a) Manual. (b) Atlas and appearance. (c) Atlas and intensity.

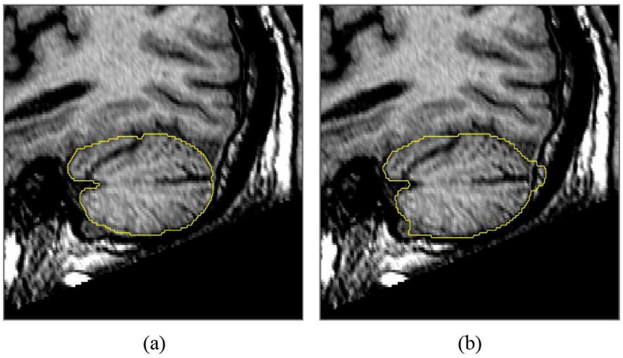
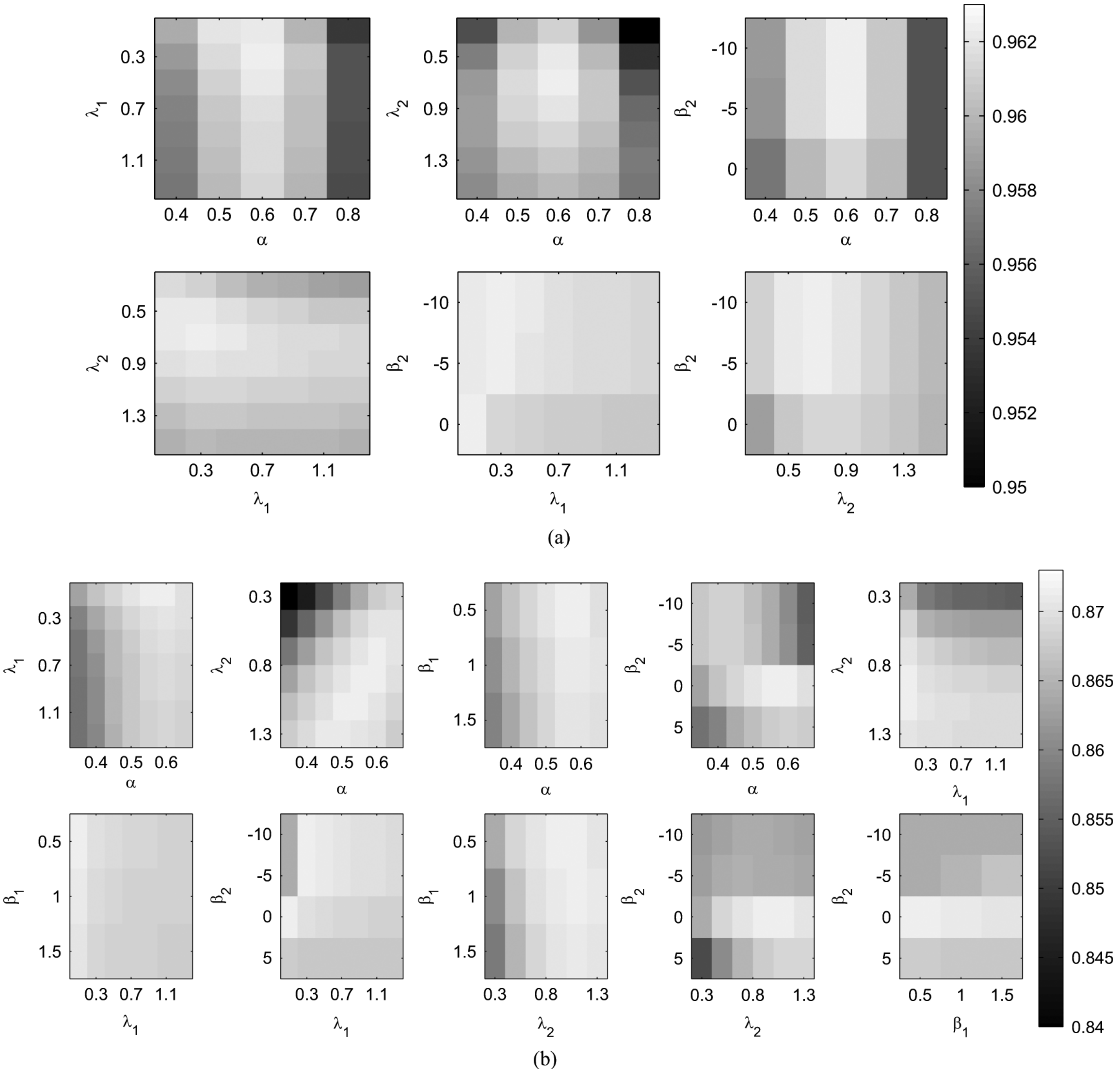


Fig. 8. Sagittal image showing an over-segmentation of the I-CRBL set caused by the spatial model that is not compensated for by the appearance model. (a) Manual segmentation. (b) Atlas- and appearance-based segmentation.

The results of I-CRBL showed some cases of over-segmentation at the posterior border with the skull, caused by registration errors in this area. The proposed method is unable to compensate for these errors as high-intensity voxels in the fatty parts of the skull were considered to be foreground by the appearance model. An extreme example can be seen in Fig. 8.

When segmenting the I-CRBL set with a smaller nine-subject training set we obtained DSI and RV scores of $0.952 \pm 0.009\,[0.927; 0.963]$ and $0.007 \pm 0.038\,[-0.110; 0.092]$. The results obtained with the 17-subject training set listed in Table II were 0.954 ± 0.008 $[0.925;\ 0.963]$ and $0.003 \pm$



(a)

(b)

Fig. 9. Average Dice similarity index computed over all target images as a function of two parameters. The other three parameters were set to the values that gave the highest average similarity index. For I-CRBL only one value of β_1 was considered during parameter learning. (a) I-CRBL. (b) I-HC.

0.039 $[-0.119;\ 0.099]$. These differences in DSI and RV were statistically significant at p-values of < 0.001 and 0.002, albeit very small. For I-HC a smaller training set resulted in a DSI of 0.864 ± 0.026 $[0.749\,;0.898]$ versus 0.870 ± 0.017 $[0.829;\ 0.899]$ and an RV of 0.031 ± 0.087 $[-0.110;\ 0.231]$ versus 0.031 ± 0.092 $[-0.122;\ 0.244]$. The difference in DSI was statistically significant at $p < 0.001$, but the difference in RV was not $(p = 0.7)$.

Fig. 9 shows the average DSI as a function of two model parameters while keeping the other three at the optimal values. For I-CRBL these optimal parameter settings were $\alpha = 0.6$, $\lambda_1 = 0.3$, $\lambda_2 = 0.7$, $\beta_1 = 1.0$, $\beta_2 = -10$. Most parameter settings around these values produced a score of 0.96 or higher, which is close to the accuracy of 0.962 obtained in the leave-one-out experiments. Since the parameter learning for the cerebellum was performed using DSI scores computed over the entire structure, this is a little higher than the score listed in Table I, which was computed over the left and right sides separately. The average maximum DSI for the I-CRBL set was 0.965.

The optimal settings for the I-HC set were $\alpha = 0.55$, $\lambda_1 = 0.1$, $\lambda_2 = 0.9$, $\beta_1 = 0.5$, $\beta_2 = 0$. Although the results are not as stable as for I-CRBL, there are several settings that give a DSI of at least 0.87, which was the score obtained in Table I. For the I-HC set the average maximum SI was 0.878.

IV. Discussion and Conclusion

The work presented in this paper demonstrates that atlas- and appearance-based models can produce robust and accurate segmentations of brain structures with both simple and complex intensity distributions. The appearance model has an increased capability to recognize neighboring background structures with overlapping intensity distributions. As a result, the proposed method can handle structures as different in shape and appearance as the hippocampus and the cerebellum. The improved separation of foreground and background also makes the method more robust to registration errors. This increases its potential for application to large-scale brain MRI studies compared to atlas- and intensity-based methods like [5]–[8].

The overall segmentation accuracy of the atlas- and appearance-based method with respect to manual labelings is good. The DSI and JSI scores on the three validation sets considered exceeded 0.85 and 0.75. The mean distance error was of the order of the voxel size or smaller. This is comparable to the results of alternative methods reported in the literature [8]–[10], [12], [13], [16], [19], [30]. The atlas and appearance-, and atlas-based segmentations did show some large deviations from the manual labellings, especially in the I-CRBL set. This was caused by inclusion of large fissures which were left out of the manual segmentation, while smaller fissures were included by the observer. As this is an error that is very particular to the cerebellum, we expect improved $D_{\max}$ scores when the method is applied to other structures.

The volume estimates derived from the proposed segmentation method showed little to no bias. The standard deviation of the volume measurements were 4% and 10% for the cerebellum and hippocampus segmentations, respectively. The scatter plot of the automated and manual cerebellar volumes showed no distinct volume-dependent biases. However, for the hippocampal volume measurements there is a tendency to underestimate large, and overestimate small volumes. This bias is likely to be caused by the multi-atlas registration: the atlas-based segmentation without appearance model has a stronger bias towards an average volume. The volumetric ICC was 0.912 for the cerebellum and 0.633 for the hippocampus in set I. The II-HC showed a higher ICC of 0.797, although its RV estimates were quite comparable to that of the I-HC set. This improvement might partly be explained by the larger range of hippocampal volumes in set II.

The atlas- and appearance-based method showed increased robustness to large registration errors. The appearance model can correct registration errors of the hippocampus when they cross over to gray matter areas like the enthorinal cortex or parahippocampal gyrus. The atlas- and intensity-based model cannot distinguish these regions from foreground. Large misregistrations occurred in about 90% of the cases in both I-HC and II-HC. This is comparable to the error rate found in an experiment on the entire cohort from which the II-HC set was taken [7]. Inclusion of an appearance model would decrease this error rate.

In images where the atlas-based registration did not show any large errors, the proposed method performed better than the strictly atlas-based method as it corrects small registration errors. However, in these cases the atlas- and intensity-based method gave comparable results. As long as the spatial model does not venture into the gray matter outside the hippocampus, our experiments suggest that intensity information alone is sufficient to improve the results. Since 90% of the cases had an accurate spatial model, the increased robustness of the appearance model had limited impact on the mean performance scores. For I-HC the spatial model was of such high quality that the atlas-based method performs almost comparably to the methods with additional components.

The segmentation results depend on the five free parameters included in the model. Our experiments showed that it is possible to learn a set of values from the training data that give results close to the maximum possible accuracy. This shows that the performance of the different model components on the training set is sufficiently representative for their performance on the target image. Furthermore, the accuracy remains relatively stable when the parameters were varied. This suggests that a less expensive parameter learning procedure could also give good results.

A limitation of the atlas- and appearance-based method is its dependency on training data that is sufficiently representative of the unlabeled target image. We have shown that the appearance model can model two different MR sequences, but it needs specific training data to accurately perform the classification and feature selection. Also the parameter learning procedure depends on the training data to estimate the quality of the model components. The proposed method is therefore primarily suitable for accurate and automated segmentation of images obtained in large neuroimaging studies where it pays off to invest in a specific training set. However, the experiment with a smaller training set suggests that you would not need the amount of training images available for this work.

The computational costs of the proposed method are high. However, the atlas registrations required for the spatial model can be parallelized. Moreover, the extra effort results in a large gain in accuracy and robustness compared to a single atlas registration [19]. The knn classifier is relatively expensive to apply to a target image, especially compared to classifiers like AdaBoost which are very fast once they have been trained [13]. We chose this method because it is flexible and easy to implement, but it could be substituted for a faster classifier, as long as it produces a probabilistic output.

The major advantage of graph cuts optimization, its ability to find a global optimum in finite time, can only be achieved for binary segmentation problems. To solve the multi-label problem of segmenting the left and right cerebellum, we had to resort to a two-step approach that used the atlas to separate the sides. Alternatively, multi-label segmentation can be performed using combinatorial optimization methods like Fast PD [31] or alpha expansions [18]. These techniques are not guaranteed to find a global optimum, but do appear to give robust results in practice.

Contrary to most previously published appearance-based segmentation methods we did not include spatial location as a feature [11], [13], [14], [16]. In this way no spatial information was contained in the appearance model. Fig. 5 and the high quality of the segmentations suggests that an appearance model without any spatial features is adequate, at least for the structures considered in this work. Jointly modeling appearance and location could help to prevent errors like the cerebellum over-segmentation shown in Fig. 8. On the other hand much of the spatial variation within the cerebellum occurs at a relatively small scale. As a result, it would be very hard to align the training images accurately enough to capture these patterns in the classifier's training set.

As we had strong spatial and appearance models, we decided to include a relatively simple interaction potential. As shown in [20], [21], the DRF framework can easily be extended to incorporate a more complex interaction potential based on a logistic classifier that gives individual weights to all feature differences. In essence, this is a classifier in its own right that labels voxel combinations instead of individual voxels. On the other hand, this model has more parameters which would require a more complex parameter learning strategy.

In conclusion, we have presented a brain structure segmentation method based on atlas registration and multi-feature classification. Because of the classifier's ability to model appearance it can segment structures with both complex and simple intensity distributions. Its accuracy with respect to manual segmentations is good, and comparable or better than existing segmentation methods. Furthermore, the appearance component makes the method more robust to large misregistration compared to atlas- and intensity-based methods.

References

[1] M. P. Laakso, K. Partanen, P. Riekkinen, M. Lehtovirta, E. L. Helkala, M. Hallikainen, T. Hanninen, P. Vainio, and H. Soininen, "Hippocampal volumes in Alzheimer's disease, Parkinson's disease with and without dementia, and in vascular dementia: An MRI study," *Neurology*, vol. 46, no. 3, pp. 678–681, 1996.

[2] C. R. Jack, R. C. Petersen, Y. C. Xu, S. C. Waring, P. C. O'Brien, E. G. Tangalos, G. E. Smith, R. J. Ivnik, and E. Kokmen, "Medial temporal atrophy on MRI in normal aging and very mild Alzheimer's disease," *Neurology*, vol. 49, no. 3, pp. 786–794, 1997.

[3] I. C. Wright, S. Rabe-Hesketh, P. W. Woodruff, A. S. David, R. M. Murray, and E. T. Bullmore, "Meta-analysis of regional brain volumes in schizophrenia," *Am. J. Psychiatry*, vol. 157, no. 1, pp. 16–25, 2000.

[4] B. Fischl, D. H. 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. M. Dale, "Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain," *Neuron*, vol. 33, no. 3, pp. 341–355, 2002.

[5] J. Zhou and J. C. Rajapakse, "Segmentation of subcortical brain structures using fuzzy templates," *NeuroImage*, vol. 28, no. 4, pp. 915–924, 2005.

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

[7] F. Van der Lijn, T. den Heijer, M. M. B. Breteler, and W. J. Niessen, "Hippocampus segmentation in MR images using atlas registration, voxel classification, and graph cuts," *NeuroImage*, vol. 43, no. 4, pp. 708–720, 2008.

[8] M. Chupin, A. Hammers, R. S. N. Liu, O. Colliot, J. Burdett, E. Bardinet, J. S. Duncan, L. Garnero, and L. Lemieux, "Automatic segmentation of the hippocampus and the amygdala driven by hybrid constraints: Method and validation," *NeuroImage*, vol. 46, no. 3, pp. 749–761, 2009.

[9] R. Wolz, P. Aljabar, J. V. Hajnal, A. Hammers, and D. Rueckert, Alzheimer's Disease Neuroimaging Initiative, "LEAP: Learning embeddings for atlas propagation," *NeuroImage*, vol. 49, no. 2, pp. 1316–1325, 2010.

[10] X. Han and B. Fischl, "Atlas renormalization for improved brain MR image segmentation across scanner platforms," *IEEE Trans. Med. Imag.*, vol. 26, no. 4, pp. 479–486, Apr. 2007.

[11] Y. Arzhaeva, E. van Rikxoort, B. van Ginneken, T. Heimann, M. Styner, and B. Van Ginneken, "Automated segmentation of caudate nucleus in MR brain images with voxel classification," in *3D Segmentation In The Clinic: A Grand Challenge*. New York: Springer, 2007, pp. 65–72.

[12] S. Powell, V. A. Magnotta, H. Johnson, V. K. Jammalamadaka, R. Pierson, and N. C. Andreasen, "Registration and machine learning-based automated segmentation of subcortical and cerebellar brain structures," *NeuroImage*, vol. 39, no. 1, pp. 238–247, 2008.

[13] J. H. Morra, Z. Tu, L. G. Apostolova, A. E. Green, C. Avedissian, S. K. Madsen, N. Parikshak, X. Hua, A. W. Toga, C. R. Jack, M. W. Weiner, P. M. Thompson, and A. D. N. Initiative, "Validation of a fully automated 3D hippocampal segmentation method using subjects with Alzheimer's disease mild cognitive impairment, and elderly controls," *NeuroImage*, vol. 43, no. 1, pp. 59–68, 2008.

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

[15] F. Ven der Lijn, M. de Bruijne, Y. Y. Hoogendam, S. Klein, K. Hameeteman, M. M. B. Breteler, and W. J. Niessen, "Cerebellum segmentation in MRI using atlas registration and local multi-scale image descriptors," in *Proc. IEEE Int. Symp. Biomed. Imag.: Macro to Nano*, 2009, pp. 221–224.

[16] J. H. Morra, Z. Tu, L. G. Apostolova, A. E. Green, A. W. Toga, and P. M. Thompson, "Comparison of Adaboost and support vector machines for detecting Alzheimer's disease through automated hippocampal segmentation," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 30–43, Jan. 2010.

[17] D. Greig, B. Porteous, and A. Seheult, "Exact maximum a posteriori estimation for binary images," *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, vol. 51, no. 2, pp. 271–279, 1989.

[18] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.

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

[20] S. Kumar and M. Hebert, "Discriminative random fields: A discriminative framework for contextual interaction in classification," *Proc. ICCV*, pp. 1150–1159, 2003.

[21] S. Kumar and M. Hebert, "Discriminative random fields," *Int. J. Comput. Vis.*, vol. 68, no. 2, pp. 179–201, 2006.

[22] J. Kittler and F. M. Alkoot, "Moderating k-NN classifiers," *Pattern Anal. Appl.*, vol. 5, no. 3, pp. 326–332, 2002.

[23] S. Arya, D. Mount, N. Netanyahu, R. Silverman, and A. Wu, "An optimal algorithm for approximate nearest neighbor searching in fixed dimensions," *J. ACM*, vol. 45, no. 6, pp. 891–923, 1998.

[24] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. W. Pluim, "Elastix: A toolbox for intensity-based medical image registration," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 196–205, Jan. 2010.

[25] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient N-D image segmentation," *Int. J. Comput. Vis.*, vol. 70, no. 2, pp. 109–131, 2006.

[26] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 147–159, Feb. 2004.

[27] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1124–1137, Sep. 2004.

[28] J. G. Sled, A. P. Zijdenbos, and A. C. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in MRI data," *IEEE Trans. Med. Imag.*, vol. 17, no. 1, pp. 87–97, Jan. 1998.

[29] K. O. McGraw and S. P. Wong, "Forming inferences about some intraclass correlation coefficients," *Psychol. Methods*, vol. 1, no. 1, pp. 30–46, 1996.

[30] K. M. Pohl, S. Bouix, M. Nakamura, T. Rohlfing, R. W. McCarley, R. Kikinis, W. E. L. Grimson, M. E. Shenton, and W. M. Wells, "A hierarchical algorithm for MR brain image parcellation," *IEEE Trans. Med. Imag.*, vol. 26, no. 9, pp. 1201–1212, Sep. 2007.

[31] N. Komodakis and G. Tziritas, "Approximate labeling via graph cuts based on linear programming," *IEEE Trans Pattern Anal Mach. Intell.*, vol. 29, no. 8, pp. 1436–1453, Aug. 2007.