SHAPE-BASED SEMI-AUTOMATIC HIPPOCAMPAL SUBFIELD SEGMENTATION WITH LEARNING-BASED BIAS REMOVAL

Hongzhi Wang1, Sandhitsu R. Das1, John Pluta1,2, Caryne Craige1, Murat Altinay1, Michael Weiner3, Susanne Mueller3, Paul A. Yushkevich1

Departments of 1Radiology and 2Neurology, University of Pennsylvania, Philadelphia, PA, USA 3Department of Veterans Affairs Medical Center, San Francisco, CA, USA

ABSTRACT
We develop a semi-automatic technique for segmentation of hippocampal subfields in T2-weighted in vivo brain MRI. The technique takes the binary segmentation of the whole hippocampus as input, and automatically labels the subfields inside the hippocampus segmentation. Shape priors for the hippocampal subfields are generated from shape-based normalization of whole hippocampi via the continuous medial representation method. To combine the shape priors with appearance features, we use a machine learning based method. The key novelty is that we treat the mistakes made by the shape priors as bias, which can be detected and corrected via learning. The main advantage of this formulation is that it significantly simplifies the learning problem by taking full advantage of current segmentations and focusing on only improving their drawbacks. Experiments show that the bias removal approach achieves significant improvement in all subfields. Our bias removal idea is general, and can be applied to improve other segmentation methods as well.

Index Terms— hippocampus, subfield, bias correction

1. INTRODUCTION
The hippocampus plays an important role in memory function [13]. Hippocampal volume has been used as an important biomarker in clinical studies, e.g. [3, 8]. However, the hippocampus is not a homogeneous brain structure. It contains several distinct subfields with different roles and susceptibility to pathology. A number of recent studies (see overview of the literature in [5]), have proposed imaging techniques and manual segmentation protocols aimed at accurately measuring hippocampal subfield volumes. Clinical utility of hippocampal subfield volumetry has been demonstrated in dementia (e.g. [7]) and other brain diseases. Hence, automatic hippocampal subfield segmentation is of great interest. [3, 10, 1] partition the boundary surface of the hippocampal ROI into regions corresponding to subfields. Since one of the most interesting subfields, the dentate gyrus, has almost no projection onto the boundary of the typical hippocampus ROI, such approaches are necessarily limited in their detail and accuracy. The most recent automatic subfield segmentation [9] uses high resolution, (0.4 × 0.4 × 0.8 mm³) T1-weighted MRI images that were acquired using specialized equipment (custom 32-channel coil) and required long scan times (5 acquisitions of 7.5 min). By contrast, our method is evaluated on T2-weighted MRI that can be acquired in a routine clinical setting with short acquisition times.

In its present form, our method takes the binary segmentation of the whole hippocampus as input, and concentrates on accurately defining subfield boundaries inside this segmentation. We build geometric priors for the subfields using shape-based normalization of whole hippocampi via the continuous medial representation (cm-rep) method [12]. To combine the shape priors with appearance features, we use machine learning. First, we generate an initial subfield segmentation using only the shape priors. The key concept of our method is that we treat the voxels labeled by the initial segmentation as bias imposed by the shape priors. In our case, this bias is mainly due to the lack of appearance features in building the shape priors. Formulated this way, it is straightforward to apply machine learning techniques to remove the bias using the missing features. In the context of machine learning, our method is related to [6], where intermediate classification results are used to improve the classifier’s performance. The main novelty of our bias removal method is that it takes full advantage of the shape priors and focuses on improving their drawbacks, which significantly simplifies the learning problem. We also propose a set of new, more efficient features for AdaBoost training. As a general concept, our bias removal idea can be used for other segmentation problems as well.

2. METHODOLOGY

2.1. Subjects, imaging and manual segmentation
We use the in vivo MRI dataset from [7], which were acquired on a Bruker Med-Spec 4T system controlled by a Siemens Trio™ console and equipped with a USA instruments eight...
channel array coil that consisted of a separate transmit coil enclosing the eight receiver coils. The hippocampus was imaged with a T2 weighted fast spin echo sequence TR/TE: 3500/19 ms, echo train length 15, 18.6 ms echo spacing, 160 flip angle, 100% oversampling in ky direction, acquisition time 5:30 min, angulated perpendicular to the long axis of the hippocampal formation. The image has $0.4 \times 0.4$ mm in plane resolution, 2 mm slice thickness, 24 interleaved slices without gap. Data with similar characteristics can be acquired on 3T scanners as well.

The manual segmentation protocol is derived from [7]; it has been expanded to include more slices and additional subfields. Each hippocampus is partitioned into anterior (head), posterior (tail) and mid-region (body), with boundaries between these regions defined by a pair of slices in the MRI image. Our work concentrates only on the hippocampal body, where manual segmentation of hippocampal subfields is feasible. The hippocampal body is divided into cornu Ammonis fields 1-3 (CA1-3), dentate gyrus (DG), and a miscellaneous label, which contains cysts, arteries, etc.

2.2. Overview of our approach

Our approach assumes that the binary segmentation of the hippocampus, as well as the slices defining the head, tail and body region, are given. The whole hippocampus segmentation is used to establish shape-based normalization across different subjects. To this end, a deformable cm-rep model [11] is fitted to each hippocampus. The model imposes a 3D coordinate system on the interior of the hippocampus, which establishes a one to one correspondence between all hippocampi. The correspondence is based on geometrical properties derived from the medial axes of the hippocampi.

Our data set contains 33 images with labeled hippocampi. To evaluate our technique, we perform a series of cross validation experiments. In each such experiment, 22 subjects are randomly selected for training, and 11 for testing. For each training hippocampus, we generate an initial shape-based labeling by mapping subfield labels from the remaining 21 training hippocampi via the cm-rep coordinate system, followed by voting. In the same way, we generate an initial shape-based labeling for the 11 test images. As described in the subsequent sections, we use the training data to learn the bias in the initial segmentation with respect to the manual segmentation. We then use the training data to learn how to correct the bias. Our method is summarized in Fig. 1.

2.3. Error analysis in the initial segmentation

From the information perspective, the initial segmentation is produced purely based on shape information. By including more informative features, we expect to have better results. From the theoretical point of view, we categorize the initial segmentation errors into: 1) consistent errors and 2) random errors. As shown in Fig. 2, in the two segmentations obtained by cm-rep modeling, some similar mistakes occur in both images. For instance, near the boundary of CA1 and DG some CA1 voxels are mislabeled as DG and some voxels from DG are mislabeled as CA1. These mistakes also appear at similar locations. For errors like this that occur across different subjects when certain conditions are met, one can effectively predict where these errors occur via capturing the causing conditions/patterns. We call this type of errors the bias imposed by the segmentation method. Besides the consistent errors, the remaining errors are noises which appear randomly and can not be reliably predicted. Although we can not do much about the random errors, it is possible to remove the consistent bias.

We propose a two-step procedure to make improvements: 1) bias detection that finds the mislabeled voxels in the initial segmentation and 2) bias correction that corrects the mislabeled voxels found by bias detection (See Fig. 1). We will substantiate these two steps in the following sections.

2.4. Bias detection as a binary classification Problem

Given the initial segmentation obtained by cm-rep modeling, our goal is to identify which voxels are mislabeled. With man-
ual segmentations, it is straightforward to formulate the bias detection problem as a binary classification problem (see Fig. 3), which we address using AdaBoost learning [4].

Fig. 3. Bias detection as a classification problem. Left to Right: manual subfield segmentation; initial segmentation; mislabeled voxels in the initial segmentation.

AdaBoost iteratively builds strong classifiers by combining complementary weak classifiers. Informally, two classifiers are complementary to each other if they do not make similar mistakes. To produce the weak classifiers, AdaBoost relies on a rich and discriminative feature pool. One common way to construct features is to use an over-complete description for each voxel and its neighborhood by convolving with a filter bank. Typically, tens of thousands of features are used. Although this over-complete description gives a good feature pool, the huge feature volume also makes learning expensive.

The motivation for using over complete description is to capture spatial correlations among local features. Our key insight is that, by combining feature-wise weak classifiers, AdaBoost algorithm also captures such correlations. Removing the redundancy can achieve better efficiency. We note that there is a compromise between the complexity of the extracted features and the complexity of AdaBoost learning. If the extracted features are complex and contain all informative patterns for the classification problem, it is straightforward for AdaBoost to pick the useful ones out. Otherwise, AdaBoost may capture the complex patterns by combining simple features. Better efficiency can be achieved via balancing the complexities between feature extraction and AdaBoost learning. We propose to use moderately complex local features and rely on AdaBoost to capture the residual spatial correlations.

We denote $A^{\Delta x, \Delta y, \Delta z}(i) = I(i_x + \Delta x, i_y + \Delta y, i_z + \Delta z) - T$ to be the appearance feature for voxel $i$ with coordinate $(i_x, i_y, i_z)$ at the relative location $(\Delta x, \Delta y, \Delta z)$. $I$ is intensity. To compensate for different intensity ranges, we normalize the intensities by the average intensity of the whole hippocampus, $T$. Note that any complex features are combinations of our appearance features. We also use the contextual feature, the subfield labels from the initial segmentation, used in [6]. We denote the contextual feature as $L^{\Delta x, \Delta y, \Delta z}(i) = s(i_x + \Delta x, i_y + \Delta y, i_z + \Delta z)$, where $s$ is the initial subfield segmentation. To include spatial information, we use $S_z(i) = i_z - \pi$, $S_y(i) = i_y - \pi$ and $S_x(i) = i_x - \pi$. Again, to compensate for different spatial ranges, $\pi$ is the average $x$ coordinates of the whole hippocampus. To enhance the spatial correlation of our features, we include the joint feature obtained from multiplying the spatial feature with the appearance and contextual features. For example, the joint features of appearance and location include $A^{\Delta x, \Delta y, \Delta z}(i)S_z(i)$, $A^{\Delta x, \Delta y, \Delta z}(i)S_y(i)$, and $A^{\Delta x, \Delta y, \Delta z}(i)S_x(i)$. Since the in slice resolution is much higher than slice thickness, we use $-6 \leq \Delta x, \Delta y \leq 6$ and $\Delta z = 0$. Overall, we use ~1300 features to describe each voxel, an order of magnitude fewer than commonly used features, e.g. by [6]. Given the response of a feature at each voxel, a weak classifier is constructed by selecting an optimal threshold to identify mislabeled voxels.

2.5. Learning-based bias correction

Bias detection outputs candidate voxels that we suspect to be mislabeled in the initial segmentation. We then need to reassign new, hopefully correct, labels to them. There are many ways to approach this problem. One simple method is to enforce the correct topology of subfields. For instance, if a voxel at the boundary of CA1 and DG is mislabeled, and its neighbor voxels are correctly labeled, because of the topology constraint, we know that the correct label for this voxel has to be either CA1 or DG. Hence, switching its label can correct it. This simple method can correct most mislabeled voxels, however it may run into trouble at 3-way or 4-way boundaries. For a more robust method, again we use a learning-based method. Given all the voxels that are mislabeled in the initial segmentation, we train classifiers to map them to the correct labels. This is a multi-class classification problem. We follow the common practice and train a binary classifier for each label to separate it from others. For this task, we use AdaBoost training with the same set of features as described above. Since we only use the mislabeled data for training, the learning cost is much less than that using the entire training data. For instance, our cm-rep based method only results in 14% mislabeled voxels (see section 3 for details). Hence, our learning cost for bias correction is only 14% compared to using the entire data. Moreover, by not taking the correctly labeled voxels into consideration, relabeling the mislabeled voxels is a simpler problem than relabeling the whole hippocampus, which simplifies the learning step as well. After bias detection, we reevaluate each detected mislabeled voxel by each classifier and assign the label whose corresponding classifier gives the highest score to the voxel. Again, since bias correction is only for detected candidate mislabeled voxels, the computational cost is much less than re-evaluate the whole hippocampus.

3. EXPERIMENTS

The results are the average of 10 cross-validation experiments. On average, the whole hippocampal body contains 1670 voxels. The initial shape-based segmentation produces only 236 mislabeled voxels. Our bias detection achieved precision of 41% at the recall of 45%. Bias correction correctly classified 63% of the detected mislabeled voxels. Overall,
our bias removal process resulted in 21% fewer mislabeled voxels (185 mislabeled voxels) in the final segmentations. Fig. 4 shows example segmentation results. Table 1 reports Dice overlaps between the initial and final segmentation results and the manual segmentations. Dice overlaps for the large subfields are high, but the reader should keep in mind that the whole hippocampus boundary is fixed in these experiments. Hence, the reported Dice overlaps must be interpreted differently than in other manual and automatic segmentation papers. As a shape based method, cm-rep works much better on large subfields, CA1 and DG, than on small subfields, CA2-3. Our bias correction achieves consistent improvements for all subfields and even more improvements for small ones. Table 1 also gives the average boundary distance with respect to the manual segmentation for each subfield. The boundary errors are measured by the point-to-mesh distance.

<table>
<thead>
<tr>
<th>method</th>
<th>CA1</th>
<th>CA2</th>
<th>CA3</th>
<th>DG</th>
</tr>
</thead>
<tbody>
<tr>
<td>initial(Dice)</td>
<td>0.919</td>
<td>0.539</td>
<td>0.686</td>
<td>0.834</td>
</tr>
<tr>
<td>final(Dice)</td>
<td><strong>0.935</strong></td>
<td><strong>0.580</strong></td>
<td><strong>0.759</strong></td>
<td><strong>0.875</strong></td>
</tr>
<tr>
<td>initial(boundary)</td>
<td>0.264</td>
<td>0.485</td>
<td>0.420</td>
<td>0.333</td>
</tr>
<tr>
<td>final(boundary)</td>
<td><strong>0.220</strong></td>
<td><strong>0.432</strong></td>
<td><strong>0.291</strong></td>
<td><strong>0.244</strong></td>
</tr>
</tbody>
</table>

4. DISCUSSIONS

Our method requires binary segmentations of whole hippocampi for cm-rep modeling. Since automatic binary hippocampus segmentation is a relatively well studied topic, e.g. [2, 6], we are currently working on fully automatic subfield segmentation via replacing the human interaction with automatic binary hippocampus segmentation. Since the concept of improving existing segmentation methods by bias removal is general, in future work we will apply it for segmentation problems other than hippocampus subfield segmentation.

5. REFERENCES


