MAPPING HIPPOCAMPAL DEGENERATION IN 400 SUBJECTS WITH A NOVEL AUTOMATED SEGMENTATION APPROACH

Jonathan H. Morra[1], Zhuowen Tu[1], Liana G. Apostolova[2], Amity E. Green[2], Christina Avedissian[1], Sarah K. Madsen[1], Neelroop Parikshak[1], Xue Hua[1], Arthur W. Toga[1], Clifford R. Jack Jr.[3], Norbert Schuff[4], Michael W. Weiner[4,5], and Paul M. Thompson[1]

[1]Laboratory of NeuroImaging, and [2]Alzheimer's Disease Research Center, UCLA School of Medicine, Los Angeles, CA

[3]Mayo Clinic College of Medicine, Rochester, MN

[4]Dept. Radiology, and [5]Dept. Medicine and Psychiatry, UC San Francisco, San Francisco, CA

ABSTRACT

We automatically segmented the hippocampus in 400 brain MRI scans from the Alzheimer's Disease Neuroimaging Initiative, using AdaBoost in conjunction with a novel model, the Auto Context Model. Our classifier, trained on 21 hand-labeled segmentations, created binary maps of the hippocampus for: 100 subjects with Alzheimer's disease (AD), 200 with mild cognitive impairment (MCI), and 100 elderly controls (mean age: 75.84; SD: 6.64). Hippocampal traces were converted to parametric surface meshes; a radial atrophy mapping technique was used to compute average surface models and local statistics of atrophy. These maps visualized correlations between regional atrophy and diagnosis (MCI versus controls: $p = 0.008$; MCI versus AD: $p = 0.001$), mini-mental state exam scores, and clinical dementia rating scores (CDR; all $p < 0.0001$, corrected). We gradually reduced sample sizes and used false discovery rate curves to determine that 40 subjects were sufficient to detect significant correlations between atrophy and CDR scores; 304 subjects were sufficient to distinguish MCI from AD.

1. INTRODUCTION

Alzheimer's disease (AD), the most common type of dementia, is associated with the pathological accumulation of amyloid plaques and neurofibrillary tangles in the brain. It first affects memory systems, progressing to involve language, affect, executive function, and all aspects of behavior. A major therapeutic goal is to assess whether treatments delay or resist disease progression before widespread cortical and subcortical damage occurs. For this, sensitive neuroimaging measures have been sought to quantify structural brain changes in early AD, which are automated enough for large-scale studies of disease and the factors that affect it.

Using MRI at millimeter resolution, subtle hippocampal shape changes may be resolved. However, isolating the hippocampus in a large number of MRI scans is time-consuming, and most studies still rely on manual tracing guided by expert knowledge of the location and shape of each region of interest (ROI). To accelerate epidemiological studies and clinical trials, some automated systems have been proposed for hippocampal segmentation [1, 2, 3], but none is yet widely used.

Pattern recognition techniques offer a range of promising algorithms for automated subcortical segmentation. Generally, pattern recognition (or machine learning) methods combine different cues to assign a probability to a specific outcome. In image segmentation, image cues are pooled to assign a specific probability to each image voxel, denoting the chance it is part of an ROI (e.g., the hippocampus) or not. In pattern recognition, cues are usually referred to as features, and different algorithms combine these features in different ways. When using pattern recognition approaches, it is standard practice to divide a dataset into two non-overlapping classes, for training and testing. The training set is used to learn the patterns (e.g., estimate a function or decision rule to classify voxels); the testing set is used to validate how well new datasets can be classified, based on the learned patterns.

Since medical images are complex, many possible features may be created to represent each voxel. Given the large number of voxels in an MRI scan, computing and storing this amount of data may become unmanageable. To overcome this problem we use a variant of a machine learning algorithm called AdaBoost [4] inside a new pattern recognition algorithm we call the auto context model (ACM). ACM may be used with any classification technique, but here we use it with AdaBoost, which has previously been found to be effective for subcortical segmentation in smaller samples of subjects [5].

This paper has two goals. First, we present an image analysis technique that may be applied to explore disease effects in large databases of anatomical images, with high throughput and automation (e.g., $N = 400$ subjects examined here). To do this, we combined our automated hippocampal segmentation method with a statistical mapping approach based on parametric surface meshes [6, 2, 7]. Second, we aimed to answer specific biological questions regarding AD. We created 3D spatial maps to reveal systematic patterns of hippocampal differences between large groups of AD, MCI, and healthy elderly subjects, and factors that affect degeneration. To confirm the clinical relevance of these anatomical measures, we correlated hippocampal atrophy with several widely used measures of brain function (mini-mental state exam and clinical dementia rating scores). To evaluate the statistical power of our mapping methods and provide practical information for users of this technique, we gradually reduced the sample size to determine how many subjects would be required, in future studies, to detect associations between hippocampal atrophy and two of our clinical measures.

2. METHODS

2.1. Subjects

We analyzed 400 brain MRI scans from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [8], a large multi-site longitudinal neuroimaging study of 800 adults, aged 55 to 90, including 200 elderly controls, 400 subjects with mild cognitive impairment, and 200 patients with AD. As part of a thorough clinical/cognitive assessment at the time of scan acquisition, each subject's mini-mental state examination (MMSE) score, and global and "sum-of-boxes" clinical dementia ratings (gCDR and sobCDR), [9] were assessed.

MMSE scores lower than 24 (out of 30) usually indicate dementia. The gCDR scores are discrete values of 0, 0.5, 1, 2, and 3, indicating no dementia, very mild, mild, moderate, or severe dementia. The sobCDR scores run from 0 to 18 in 0.5 intervals, (0 is no dementia; 18, very severe dementia). Our goal was to create statistical maps correlating hippocampal morphology and different covariates of interest, including diagnosis (normal, MCI, AD), MMSE, gCDR, and sobCDR.

2.2. Training and Testing Set Descriptions

When using a pattern recognition approach to identify structures in images, two non-overlapping sets of images must be defined, for training and testing. The training set consists of a small representative sample of brain images, manually traced by experts. The testing set is a separate group of brain images that are to be segmented by the algorithm, but have not been used for training the algorithm. Our training set consisted of 21 brain images (7 AD, 7 MCI, and 7 controls) and our testing set consisted of 400 brain images (100 AD, 200 MCI, and 100 controls). We used 400 testing brains, in 3 diagnostic groups whose size matched the expected proportions of the final ADNI sample (data collection is in progress). The three diagnostic groups were age- and gender-matched as closely as possible, as shown in Table 1.

	Age (years)	MMSE	gCDR	sobCDR
N	76.62 (4.83)	29.14 (0.86)	0.00 (0.00)	0.02 (0.09)
M	75.45 (7.03)	26.94 (1.86)	0.50 (0.00)*	1.48 (0.84)*
A	75.86 (7.25)	23.41 (1.86)	0.78 (0.25)*	4.48 (1.56)*

Table 1. Demographic data are shown for age, MMSE, gCDR, and sobCDR, with standard deviations in parentheses for the three diagnostic groups: normal (N), MCI (M) and AD (A). Information is shown for the test group, as that is the group used for the statistical maps. * $p < 0.01$

2.3. Feature Selection

The first step in designing a machine learning algorithm is to select relevant image features. To make our algorithm efficient, these features should be informative and quick to calculate. We therefore designed our feature pool to consist of image intensity, combinations of x, y, and z positions, image curvatures, image gradients, tissue classification maps of gray matter, white matter, and CSF, mean filters, standard deviation filters, and Haar filters of sizes varying from 1x1x1 to 7x7x7. We also linearly registered all brains to a standard template to devise a basic shape prior to capture the global shape of the hippocampus, defined as the pointwise summation of all the training masks. Also, because the subcortical structures we are investigating are contiguous geometric shapes, the classification of neighboring voxels should influence each other. Therefore, we included curvature, gradient, mean, standard deviation, and Haar filters of the shape prior as well. Overall, we used approximately 13,000 features.

2.4. Segmentation Description

AdaBoost [4] has been shown to be an effective segmentation technique in brain MRIs in previous studies [5]. AdaBoost is a weighted voting algorithm, which combines "weak learners" into a "strong learner". A weak learner is any pattern recognition algorithm that guesses correctly greater than half of the time.

Our extension of AdaBoost, the Auto Context Model, is a way of incrementally approaching the correct posterior distribution for a given dataset. During each iteration of ACM, the posterior distribution obtained is fed back in as input for the next iteration. In the context of AdaBoost, that means that it becomes a new feature for AdaBoost to use as a weak learner. However, the true power of ACM is realized when using neighborhood features of the posterior distribution as features as well due to the regularity of shape patterns observed in medical imaging. One way to view the shape prior is as an initial estimate of the posterior distribution, which is updated during each iteration of ACM. After about four iterations of ACM, the posterior distribution does not change, indicating that new image-based clues are not being discovered. ACM is also very quick, taking less than 1 minute to segment an ROI from a brain MRI on a desktop computer. Figure 1 describes both AdaBoost and ACM.

Given: N labeled training examples (x_i, y_i) with $y_i \in \{-1, 1\}$ and $x_i \in \mathbb{R}$ and a prior distribution $P_1(i)$
For $s = 1, \ldots, S$

- Create an initial uniform distribution of weights $D_1(i)$ over the examples.
- For $t = 1, \ldots, T$:
  - $\epsilon_j = \sum_{i=1}^{N} D_t(i) \left[y_i \neq h_j(x_i)\right]$
  - $h_t = \arg\min_{h_j \in H} \epsilon_j$
  - Set $\alpha_t = \frac{1}{2} \log\left((1 - \epsilon_t)/\epsilon_t\right)$.
  - Set $D_{t+1}(i) = D_t(i) \exp\left(-\alpha_t y_i h_t(x_i)\right)/Z_t$
    $Z_t = 2\sqrt{\epsilon_t(1 - \epsilon_t)}$, a normalization factor
- Calculate $P_s(x) = 1/\left(\exp\left(-f(x)\right) + 1\right)$
  $f(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$
- Update those weak learners based on $P_s(i)$

Fig. 1. The AdaBoost algorithm inside of the Auto Context Model. $\mathbf{1}$ is an indicator function. Weak learners h_t are assumed to output $\{-1, +1\}$

2.5. Linking Shape and Disease Factors

After all hippocampal segmentations had been performed, we correlated hippocampal shape with different disease-related factors using surface-based statistical maps. To accomplish this, 3D parametric surface models were constructed from each segmentation, and geometrically averaged across subjects within each diagnostic group. This results in 3D average surface maps for each diagnostic group, and statistical maps relating morphology to diagnosis and clinical scores. We employed a surface averaging approach to establish pointwise correspondence for subcortical surfaces.

To create a measure of 'radial size' for each subject's hippocampus, first a medial curve was computed threading through the hippocampus, and the distance from each surface point to this curve was calculated, providing a measure that is sensitive to local atrophy. Regressions were performed to assign a *p*-value to each point on the surface in order to link radial size to different covariates of interest. Finally the *p*-maps are presented as color-coded average subcortical shapes.

2.6. False Discovery Rate

To assess our method's power to establish linkages between morphology and disease, we created cumulative distribution function (CDF) plots of the p-values in our subcortical maps. These CDF plots (or Q-Q plots) are involved in the false discovery rate (FDR) method to assign overall significance values to statistical maps [10].

A CDF of the p-values may be used to assess how well a method can capture a known relationship between anatomy and disease, or to discover new relationships. In a plot of the observed p-values versus those expected under the null hypothesis (of no correlation), the line $y = x$ represents the null distribution. Large upward inflections of this line typically represent significant relationships, as reflected in the p-maps. Specifically, the intersection of the CDFs with the $y = 20x$ line represents the highest p-value for which at most there are 5% false positives (the false discovery rate). This intersection point, if it occurs, is called the q-value. If there is no such intersection point (other than the origin), there is no evidence to reject the null hypothesis.

3. RESULTS

3.1. Error Metrics

Since we only have ground truth data for the training set, we have to use a leave-one-out analysis to artificially create testing data with ground truth. In this analysis, 21 models were trained, each one ignoring one of the training brains. Each model was then tested on the brain that it ignored, which secures the independence between testing and training samples that is necessary for validation. We defined a variety of error metrics using the following definitions: A, the manually segmented ROI, B, the automatically segmented ROI, and $d(a, b)$, the Euclidean distance between points a and b.

- $Precision = \frac{A \cap B}{B}$
- $Recall = \frac{A \cap B}{A}$
- $RelativeOverlap = \frac{A \cap B}{A \cup B}$
- $SimilarityIndex = \frac{A \cap B}{\frac{A+B}{2}}$
- $Mean = avg_{a \in A}(min_{b \in B}(d(a, b)))$
- $H_1 = max_{a \in A}(min_{b \in B}(d(a, b)))$
- $H_2 = max_{b \in B}(min_{a \in A}(d(b, a)))$
- $Hausdorff = \frac{H_1 + H_2}{2}$

	Prec.	Rec.	R.O.	S.I.	Haus.	Mean
L Train	0.869	0.916	0.805	0.892	2.38	0.0033
R Train	0.860	0.918	0.798	0.887	2.94	0.0040
L LOO.	0.827	0.857	0.717	0.834	3.37	0.0053
R LOO.	0.801	0.841	0.693	0.816	4.56	0.0097

Table 2. Accuracy (Acc.), precision (Prec.), recall (Rec.), relative overlap (R.O.), similarity index (S.I.), Hausdorff distance (Haus.; in mm), and mean distance (in mm) are reported for the left (L) and right (R) training set, and the leave-one-out (LOO.) analysis (N = 21 for both).

As shown in Table 2, segmentation accuracy was high on the training set, and the error metrics deteriorated only slightly when performance was assessed on the leave-one-out analysis. This shows that the models are not memorizing the training examples, and should generalize well to the testing set.

3.2. Diagnostic and Clinical Measurement Differences

After segmenting the hippocampus, we made significance maps (p-maps) for diagnoses and clinical measurements. Figure 2 shows the p-maps for each pairwise diagnostic comparison, and correlations between atrophy and MMSE scores, gCDR, and sobCDR scores, as the covariates. The level of atrophy is strongly associated with diagnosis (with greatest effects for the AD v. normal comparison), MMSE scores, and CDR scores linkages. The overall significance of these mapping results was confirmed by permutation testing (Table 3).

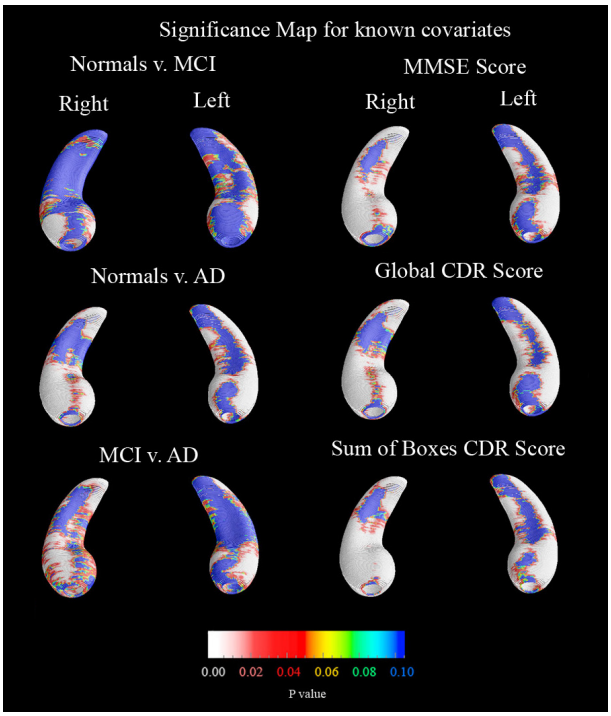

Fig. 2. Significance maps (p-value maps) show strong associations between hippocampal shape (local volumetric atrophy) and diagnosis (*left columns*) and cognitive and clinical scores (*right columns*). All 6 maps show strong statistical correlations that were confirmed in permutation tests. For the diagnostic comparisons, we used the number of subjects in each of the two groups being compared, for the clinical measurement groups we used all 400 subjects

	Left	Right
Normal v. MCI	0.00784	0.00884
Normal v. AD	0.0001	0.00011
MCI v. AD	0.00211	0.000415
MMSE	0.0001	0.0001
gCDR	0.0001	0.0001
sobCDR	0.0001	0.0001

Table 3. This table shows the permutation-corrected p-values for all maps shown in this paper. Permutation testing involved finding a p-value, in 100,000 random assignments, for the proportion of supra-threshold surface points in the p-maps. All results are significant by permutation testing.

3.3. Reducing the N

Another avenue that we explored was how many subjects were necessary to detect a statistically significant linkage between diagnosis or clinical measurements and morphology. To investigate this, we

randomly threw out subjects from our initial sample of 400, yielding groups of size 304, 200, 104, 40, and 24. These groups preserved the 1:2:1 relationship between normal, MCI, and AD sample sizes, and 1:1 gender ratio, so they are all multiples of 8. For each different N, a random number was used to throw out samples; therefore, smaller samples are not necessarily subsets of the larger ones.

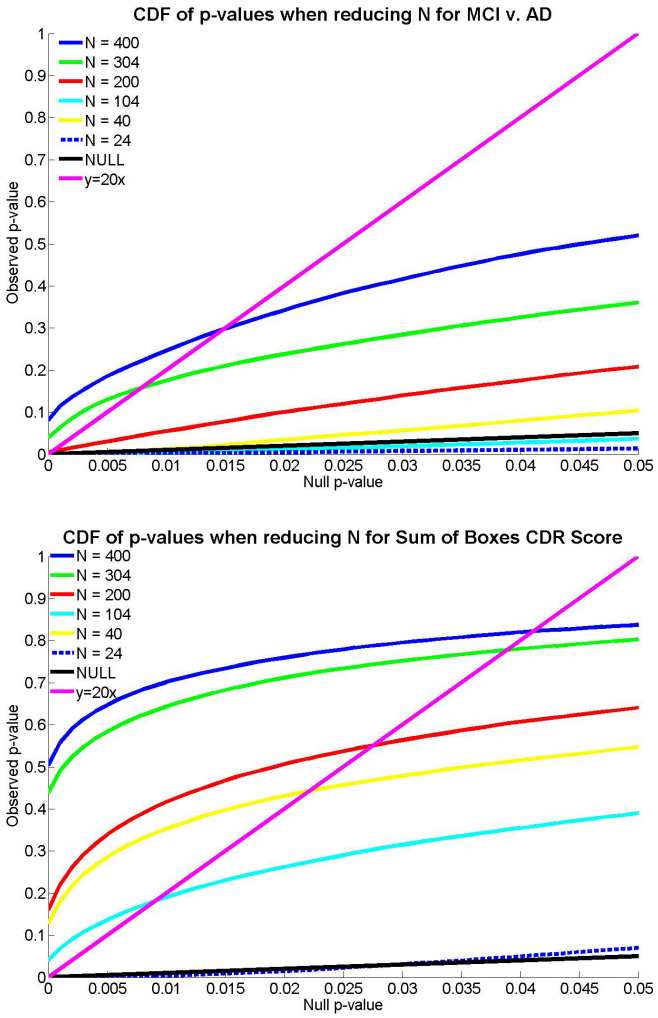

Fig. 3. CDFs of p-values measuring the effect sizes for correlations between hippocampal atrophy and two different covariates, as the sample size, N, decreases. There is not a monotonically decreasing relation between sample size and the height of the CDF computed from the sample, as each sample is the result of random sampling from a population. In general however, as N decreases, the power to detect a given effect is less. To determine the minimal sample size, we find the smallest N with a non-zero q-value.

As shown in Figure 3, reducing the sample size, N, generally decreases the detected effect size; greater effect sizes are shown by CDFs with the most rapid upswings from the origin. MCI is difficult to distinguish from AD based on atrophy, requiring 304 subjects. The sobCDR measures AD progression with a continuous rather than a categorical variable, so fewer subjects (40) were necessary. In general, more extensive regions of statistical association were detected with larger samples.

3.4. Conclusion

ACM is an effective extension to AdaBoost for segmenting the hippocampus. AdaBoost selects features based on a training set of expert segmentations, so it may generalize well for segmenting other subcortical structures, such as the basal ganglia. We will further validate ACM with AdaBoost by evaluating it on a broader range of manually segmented (ground truth) data, to validate that it is indeed finding the hippocampus (and other subcortical structures).

We further validated our model by showing statistical linkages between diagnosis and clinical scores and radial atrophy. Although some of these linkages have been established before, this is one of the first projects to do so on such a large group of patients, using a fully automated approach and using surface-based maps rather than simple volumetric measures.

Finally, we showed the number of subjects necessary to observe two different effects, distinguishing MCI from AD and linking the sobCDR with atrophy. The more subtle MCI v. AD comparison required 304 subjects and the less subtle sobCDR measurement required only 40.

4. REFERENCES

[1] B. Fischl et al., "Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain," *Neurotechnique*, vol. 33, pp. 341–355, 2002.

[2] L. Wang et al., "Large deformation diffeomorphism and momentum based hippocampal shape discrimination in dementia of the Alzheimer type," *IEEE TMI*, vol. 26, no. 4, pp. 462–470, 2007.

[3] P. A. Yushkevich, J. Piven, C. Hazlett, H. Smith, G. Smith, R. Ho, S. Ho, J. Gee, and G. Gerig, "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *Neuroimage*, vol. 31, pp. 1116–1128, 2006.

[4] R.E. Schapire and Y. Freund, "Boosting the margin: A new explanation for the effectiveness of voting methods," *Annals of Statistics*, vol. 26, no. 5, pp. 1651–1686, 1998.

[5] Z. Tu, K. Narr, I. Dinov, P. Dollár, P. Thompson, and A. Toga, "Brain anatomical structure parsing by hybrid discriminative/generative models," *IEEE Transactions on Medical Imaging*, in press 2007.

[6] J. Csernansky, L. Wang, S. Joshi, J. Ratnanather, and M. Miller, "Computational anatomy and neuropsychiatric disease: Probabilistic assessment of variation and statistical inference of group difference, hemispheric asymmetry, and time-dependent change," *NeuroImage*, vol. 23, pp. S56–S68, 2004.

[7] R. Bansal, L.H. Staib, D. Xu, H. Zhu, and B.S. Petersen, "Statistical analyses of brain surfaces using Gaussian random fields on 2-D manifolds," *IEEE TMI*, vol. 26, no. 1, pp. 46–57, Jan. 2007.

[8] C.R. Jack et al., "The Alzheimer's Disease Neuroimaging Initiative (ADNI): The MR imaging protocol," For the ADNI Consortium Study, in press 2007.

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

[10] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J.R. Statist. Soc. B*, vol. 57, no. 1, pp. 289–300, 1995.