Introduction

Alzheimer’s disease (AD) is the most common form of dementia. It is a devastating disease for those who are affected and presents a major burden to carers and society. The worldwide prevalence of AD is predicted to quadruple from 26.6 million in 2006 to more than 100 million by the year 2050 (Brookmeyer et al., 2007). Genetic risk factors have been identified (Lehtovirta et al., 1996; Harold et al., 2009). A definitive diagnosis, however, requires histological examination of brain tissue. In practice, AD is diagnosed from the patient’s history and clinical presentation, while neuroimaging is used as an adjunct (Dubois et al., 2007). Much research effort is directed at developing imaging biomarkers, motivated by the desire to increase diagnostic accuracy and to enable earlier diagnoses.

The extraction of biomarkers from structural magnetic resonance (MR) images forms a major field of research. The main focus in this area is directed at measurement of cortical thickness (Evans et al., 2005) and volume measurements of brain structures. The hippocampus is one of the first structures in the brain to be affected by Alzheimer’s disease (Braak and Braak, 1991), and hippocampal volume and especially atrophy over time has been shown to correlate with disease progression, e.g. Crum et al. (1999), Jack et al. (2004). Estimates of hippocampal atrophy in longitudinal MR images can give insights into onset and progression of dementia and can serve as biomarkers helping to discriminate dementia patients from healthy subjects. Since manual determination of the volume of brain structures is time-consuming and requires careful examination of
intra-rater and inter-rater reliability, many efforts have been devoted to developing automated methods of atrophy rate measurement: Freeborough and Fox (1997) proposed the boundary shift integral (BSI) that measures atrophy from the difference of a structure’s boundaries in baseline and registered follow-up scan. SIENA is a method that quantifies atrophy from the movement of image edges between timepoints (Smith et al., 2002). In tensor-based morphometry (TBM), the Jacobian determinants obtained from non-rigidly registering a follow-up scan to its baseline are integrated to measure atrophy (Boyes et al., 2006; Leow et al., 2007). Alternatively, volume differences can be established by segmenting a structure of interest at different timepoints (Fox et al., 2000; Barnes et al., 2008; Morra et al., 2009; Schuff et al., 2009). A technique proposed by Thompson et al. (2004) that combines 3D parametric surface mapping of a structure at Baseline and follow-up with automatic segmentation has recently been applied to the measurement of hippocampal atrophy in subjects from the ADNI study (Morra et al., 2009). When measuring subtle volume changes caused by atrophy, a consistent segmentation procedure for all timepoints is crucial. Simultaneous segmentation of image sequences has been shown to increase the accuracy of atrophy measurement (Xue et al., 2006).

The majority of existing methods addresses the segmentation of single timepoints only. A method based on graph cuts (Boykov et al., 2001) and multi-atlas label propagation (Heckemann et al., 2006) has been applied successfully to the segmentation of the hippocampus and subcortical structures (van der Lijn et al., 2008; Wolz et al., 2009). In this work we extend this algorithm to the simultaneous segmentation of a series of MR images acquired from the same subject. A subject-specific probabilistic atlas of a structure of interest is generated for each baseline image. After affine registration of follow-up scans to their baseline scan, this probabilistic atlas is used as spatial prior for all timepoints. This spatial prior, together with an intensity model derived from the unseen image, provides the data field (MRF) which defines a graph on the subject-specific probabilistic atlas of a structure of interest is generated for each baseline image. After affine registration of follow-up scans to their baseline scan, this probabilistic atlas is used as spatial prior for all timepoints. This spatial prior, together with an intensity model derived from the unseen image, provides the data field (MRF) which defines a graph on the

Materials and methods

Image data

Images used in this study were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI, Mueller et al., 2005). In the ADNI study brain MR images are acquired at baseline and regular intervals from approximately 200 cognitively normal older subjects, 400 subjects with MCI, and 200 subjects with early AD. A more detailed description of the ADNI study as well as image acquisition and preprocessing steps provided in Appendix A.

Table 1

Table 1: For each group the total number of subjects, number of females, and the average change in MMSE and CDR scores are shown, along with the comparison of the clinical values at Baseline and Month 12 scan of 12.96±1.32 months do not vary significantly on t-test between the groups.

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>F</th>
<th>ΔMMSE</th>
<th>ΔCDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN</td>
<td>163(73)</td>
<td>29.08±1.03</td>
<td>-0.07±1.39</td>
<td>0.0±0.19</td>
</tr>
<tr>
<td>MCI</td>
<td>126 (63)</td>
<td>27.02±1.74</td>
<td>-0.72±2.64</td>
<td>0.5±0.20</td>
</tr>
<tr>
<td>S-MCI</td>
<td>16 (60)</td>
<td>27.25±1.71</td>
<td>-0.03±2.35</td>
<td>0.5±0.20</td>
</tr>
<tr>
<td>P-MCI</td>
<td>63 (22)</td>
<td>26.57±1.57</td>
<td>-0.97±1.95</td>
<td>0.5±0.14</td>
</tr>
<tr>
<td>P-MCI</td>
<td>40 (21)</td>
<td>26.88±1.89</td>
<td>-2.79±2.83</td>
<td>0.5±0.20</td>
</tr>
<tr>
<td>AD</td>
<td>126 (63)</td>
<td>23.48±1.85</td>
<td>-2.59±4.09</td>
<td>0.74±0.25</td>
</tr>
</tbody>
</table>

Table 2

Table 2: Subpopulation for which three timepoints were available. The number of subjects, number of females and average change in MMSE and CDR during 24 months are given for the six subject groups.

<table>
<thead>
<tr>
<th>Group</th>
<th>N(F)</th>
<th>ΔMMSE</th>
<th>ΔCDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN</td>
<td>114 (54)</td>
<td>-0.16±1.29</td>
<td>0.06±0.06</td>
</tr>
<tr>
<td>MCI</td>
<td>165 (55)</td>
<td>-2.11±3.79</td>
<td>0.10±0.32</td>
</tr>
<tr>
<td>S-MCI</td>
<td>90 (29)</td>
<td>-0.47±2.59</td>
<td>0.03±0.17</td>
</tr>
<tr>
<td>P-MCI</td>
<td>47 (16)</td>
<td>-4.02±4.04</td>
<td>0.16±0.29</td>
</tr>
<tr>
<td>P-MCI</td>
<td>28 (10)</td>
<td>-4.18±4.22</td>
<td>0.41±0.45</td>
</tr>
<tr>
<td>AD</td>
<td>83 (39)</td>
<td>-4.43±5.64</td>
<td>0.47±0.58</td>
</tr>
</tbody>
</table>

Subjects

In this work we used those subjects for whom a Baseline and Month 12 follow-up 1.5 T scan were available (n=568). For 362 subjects within this population, a Month 24 follow-up image was also available. All images where downloaded in April 2009. For 112 subjects, progression from MCI to AD has been reported during the study. We separately looked at the subject group that converted between Baseline and Month 12 follow-up (P-MCI<12) and the group that converted at any point after the Month 12 scan (P-MCI<12), as well as the group of subjects which had a stable diagnosis of MCI (S-MCI). While the ADNI study aims to follow all subjects for 36 months, for most subjects the examination for this timepoint was not available in September 2009, which means that some subjects in the S-MCI group are likely to convert to P-MCI<12 in the future. An overview of the subject groups is given in Table 1: For each group the total number of subjects, number of females, and the average MMSE and CDR scores are shown, along with the development of these clinical values over time. The mean age for all subjects of 75.3±6.6 years and the mean time passed between Baseline and Month 12 scan of 12.96±1.32 months do not vary significantly on t-test between the groups.
Hippocampus atlases

The atlases used to automatically segment the hippocampus in baseline and follow-up images are based on hippocampal label maps provided by ADNI. To define these label maps, semi-automated hippocampal volumetry was carried out using a commercially available high dimensional brain mapping tool (Medtronic Surgical Navigation Technologies, Louisville, CO), that has previously been compared to manual tracing of the hippocampus (Hsu et al., 2002). First, 22 control points were placed manually as local landmarks for the hippocampus on the individual brain MRI data: one landmark at the hippocampal head, one at the tail, and four per slice (i.e., at the superior, inferior, medial and lateral boundaries) on five equally spaced slices perpendicular to the long axis of the hippocampus. Second, fluid image transformation was used to match the individual brains to a template brain (Christensen et al., 1997). Transformed label maps were inspected and if necessary manually corrected by qualified reviewers. Empirically, we found that the resulting hippocampal delineations start anteriorly with their separation from the amygdalae; include the bulk of the hippocampal subfields CA1−4 (Lorente de Nó, 1934), the subiculum, the dentate gyrus; miss some of the medial hippocampal head at the level of the uncus; but contain most of the intralimbic gyrus, the alveus as well as much of the fimbria, and end posteriorly shortly posterior to where cella media, temporal horn, and occipital horn fuse on coronal slices.

Although other hippocampus definitions exist (e.g. Niemann et al., 2000; Hammers et al., 2003) and can be used with the proposed method, we applied the protocol used by ADNI to allow comparison with other methods.

3D image segmentation with graph cuts

Energy functions based on Markov random fields (MRF) in combination with graph cuts have been used widely for labeling problems in computer vision (Greig et al., 1989; Boykov et al., 2001; van der Lijn et al., 2008; Song et al., 2006; Khan and Shah, 2009). Segmenting an image I is usually defined as assigning a label \( f_p \in L \) to each voxel \( p \in I \). An MRF-based energy function can be formulated as

\[
E(f) = \sum_{p \in I} D_p(f_p) + \sum_{\{p,q\} \in N} V_{pq}(f_p, f_q),
\]

where \( N \) is a neighborhood of voxels and \( f \) is the labeling of \( I \) (Boykov et al., 2001). The data term \( D_p(f_p) \) measures the disagreement between a prior probabilistic model and the observed data. \( V_{pq}(f_p, f_q) \) is a smoothness term penalizing discontinuities in \( N \). The parameter \( \lambda \) governs the influence of the data and smoothness terms. In previous work (Wolz et al., 2009) we found that setting \( \lambda = 2 \) leads to robust results for hippocampal segmentation.

To optimize Eq. (1) with graph cuts, a graph \( G = (V, E) \) with a node \( v \in V \) for each voxel \( p \) is defined on image \( I \). Its edges \( e \in E \) consist of each voxel \( v \) and two terminal nodes \( F, B \) (sink and source) as well as connections between neighboring voxels. The terminals \( F \) and \( B \) represent the two labels describing foreground and background, respectively. By determining an \( F-B \) cut on \( G \), the desired segmentation can be obtained (Boydov et al., 2001).

4D image segmentation with graph cuts

The segmentation of a structure in serial images may vary between scans even if there are only small variations in the intensity (Xue et al., 2006). This is more likely near indistinct boundaries, e.g. between hippocampus and amygdala. To be more robust against intensity variations between timepoints and against differences caused by image noise, we extend the segmentation from a single image to the simultaneous segmentation of a sequence of images. This is achieved by extending the graph defined by the energy function in Eq. (1) from 3D to 4D. A 4D image \( I \) is not only defined by spatial coordinates \( x, y, z \) but also by a time coordinate \( t \). 4D images are generated for each subject by affine registration of the follow-up scans to their baseline image, establishing correspondence between voxels in 4D.

For a 4D image, a voxel \( p^{x,y,z,t} \) has a 8-neighborhood \( N \) which incorporates the two temporally adjacent voxels \( p^{x,y,z,t-1} \) and \( p^{x,y,z,t+1} \) into the standard 6-neighborhood in 3D.

The smoothness constraint thus applies both in space and time, and the segmentations at different timepoints are forced to be consistent in areas where only a small gray value difference between the images exists. The difference in the segmentation result of neighboring timepoints can then be expected to reflect intensity differences caused by atrophy and is less likely to be caused by noise in individual images.

Energy terms

We use similar energy terms to those used in van der Lijn et al. (2008) and Wolz et al. (2009): The data term \( D_p(f_p) \) is estimated from a spatial prior and a probabilistic model of the intensity of the structure of interest. The spatial prior is obtained from registering multiple atlases to the target image and interpreting the propagated label maps probabilistically. We use the intensity model estimated from the target image, as proposed in Wolz et al. (2009), instead of relying on manual training as suggested in van der Lijn et al. (2008).

The spatial prior for the baseline images is generated by our previously proposed LEAP framework (Wolz et al., 2010). In this framework, a diverse set of brain images (e.g. images from a large multi-center trial such as ADNI) and an initial set of manually labeled atlases are embedded in a low-dimensional manifold where neighborhoods represent similarity according to a chosen measure. The initial set of atlases is propagated in several steps through the manifold using multi-atlas label propagation. That way, only similar images are registered directly; this has been shown to reduce segmentation errors (Wolz et al., 2010).

After applying LEAP, \( N \) atlases have been registered to every image in the dataset. The spatial probability of observing a structure of interest (foreground) is determined for each voxel \( p^{x,y,z,t} \) from these atlases:

\[
P_A(p^{x,y,z,t}) = \frac{1}{N} \sum_{i=1}^{N} \left\{ \begin{array}{ll} 1, & f_p^{i} = f^{x,y,z,t} \\ 0, & \text{else} \end{array} \right. \]

with \( f^{x,y,z,t} \) defining the foreground label.

After affine registration of follow-up images to their baseline, the probabilistic atlas produced for the baseline image is used for all timepoints. To establish one-to-one correspondences, voxel grids of follow-up images are aligned with that of the baseline using an interpolation based on B-splines (Unser et al., 1991). Tissue loss resulting from Alzheimer’s disease can be observed as a shift of the boundaries of anatomical structures. This means that differences in the segmentations of different timepoints can be expected to lie primarily in the boundary region of structures. Since the prior probability values of the atlas are low in the boundary regions, the segmentation in these areas depends mainly on the intensity model. A consistent gray value difference between two timepoints at a particular location therefore results in a segmentation difference which will be interpreted as atrophy.

To account for global intensity differences in individual scans, intensities in the follow-up scans are matched to those in the baseline scan using linear regression. A Gaussian probability distribution is used as the intensity model \( P_I(p, f) \). It is defined from the voxels in the image sequence where the prior probability \( P_I \) of observing the structure of interest is at least 95%.
PB p

fl

spinal
tissue classes (white matter (WM), gray matter (GM) and cerebro-

Peeling each node to the source

combined to give the data term,

parameters

fi

Gaussians (MOG) model. It is de

certain voxel

p

additional parameter

timepoints) and temporal edges (between timepoints), we use an

in Song et al. (2006). To discriminate between spatial edges (within

neighboring voxels as well as image gradients, as originally proposed

with intensity

yp

is estimated from a mixture-of-

k

= 1 , 2 , 3 ( Van Leemput, 1999) and previously

generated and non-rigidly aligned probabilistic atlases γk for three
tissue classes (white matter (WM), gray matter (GM) and cerebro-

spinal fluid (CSF)):

P B ( p , f B ) = ∑ k = 1 3 γ k P ( y p | τ k ) .

(3)

The intensity and spatial contributions, P y , X ∈ F , B and P a , are
combined to give the data term, D a , that defines the edge weights
connecting each node to the source F and sink B (Boykov et al., 2001).
The smoothness term V p,q determining the weight of an edge
connecing two voxels p , q is based on intensity differences between
neighboring voxels as well as image gradients, as originally proposed
in Song et al. (2006). To discriminate between spatial edges (within
timepoints) and temporal edges (between timepoints), we use an
additional parameter αpq weighting each edge individually.

Experiments and results

The proposed 4D graph cuts method was applied to the two image
sets described in the Image data section: Set 1, consisting of 555 image
pairs at Baseline and Month 12 follow-up and Set 2, consisting of 352
image triplets at Baseline, Month 12 and Month 24.

Fig. 1 shows a typical segmentation result for Baseline and Month 12
images on a transverse section of the right hippocampus in a
subject with AD. The atrophy-related discrepancy of the strong GM-

CSF boundary is accurately captured and, more importantly, a
consistent segmentation across timepoints is produced in areas
where the hippocampus is not defined by clear boundaries.

Hippocampal atrophy after 12 and 24 months

Atrophy rates in Set 1 are shown in Table 3. All subject groups
show a statistically significant volume loss with p < 0.001 on a paired t-

mean. Atrophy rates (%) are shown along with the standard
deviations displayed for the three clinical groups (AD, MCI, controls
(CN)) as well as the different groupings of MCI subjects introduced in
the Subjects section.

Table 4 shows the average atrophy rates (%) over 24 months when
segmenting Baseline, Month 12 and Month 24 images simultaneously.

Fig. 2 shows box-and-whisker plots of atrophy rates over 12 and
24 months for normal controls, MCI converters (P-MCI), subjects
with stable MCI (S-MCI), and AD. The difference between all clinical
groups (CN, MCI, AD) is statistically significant (p < 0.001) on a two-
sample (unpaired) t-test. No significant difference was observed
between P-MCI and AD, which can be explained by the fact that
subjects in the P-MCI group later convert to the AD group and are
therefore likely to be pathomorphologically similar.

To investigate the consistency of the proposed method as well as
the influence of additional constraints when segmenting more than
two timepoints, we compared the atrophy results obtained for the
first year when segmenting two and three timepoints simultaneously.
T-tests indicate no significant difference between the means of
matched samples (p = 0.57). A Bland–Altman plot comparing both
measures is displayed in Fig. 3. The plot shows good agreement
between the two measurements with few outliers.

Table 3

Hippocampal atrophy rates (%) in 555 subjects over 12 months. Number of subjects is
given in parentheses. Mean ± std.

<table>
<thead>
<tr>
<th></th>
<th>CN (163)</th>
<th>MCI (268)</th>
<th>AD (124)</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>0.78 ± 1.77</td>
<td>2.19 ± 2.88</td>
<td>3.82 ± 2.25</td>
</tr>
<tr>
<td>l</td>
<td>0.92 ± 1.89</td>
<td>2.36 ± 2.47</td>
<td>3.96 ± 2.51</td>
</tr>
<tr>
<td>r+l</td>
<td>0.85 ± 1.59</td>
<td>2.34 ± 2.12</td>
<td>3.85 ± 1.99</td>
</tr>
<tr>
<td>S-MCI (156)</td>
<td>P-MCI (112)</td>
<td>P-MCI12 (49)</td>
<td>P-MCI12 (53)</td>
</tr>
<tr>
<td>r</td>
<td>1.68 ± 3.12</td>
<td>2.97 ± 2.28</td>
<td>3.27 ± 2.09</td>
</tr>
<tr>
<td>l</td>
<td>1.67 ± 2.23</td>
<td>3.41 ± 2.44</td>
<td>4.00 ± 2.20</td>
</tr>
<tr>
<td>r+l</td>
<td>1.72 ± 1.91</td>
<td>3.23 ± 2.10</td>
<td>3.61 ± 1.91</td>
</tr>
</tbody>
</table>

Table 4

Hippocampal atrophy rates (%) in 352 subjects over 24 months. Number of subjects is
given in parentheses. Mean ± std.

<table>
<thead>
<tr>
<th></th>
<th>CN (114)</th>
<th>MCI (157)</th>
<th>AD (81)</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>1.52 ± 2.29</td>
<td>4.36 ± 3.26</td>
<td>6.71 ± 3.27</td>
</tr>
<tr>
<td>l</td>
<td>1.80 ± 2.19</td>
<td>4.65 ± 3.49</td>
<td>6.87 ± 3.19</td>
</tr>
<tr>
<td>r+l</td>
<td>1.66 ± 2.07</td>
<td>4.50 ± 3.12</td>
<td>6.74 ± 2.89</td>
</tr>
<tr>
<td>S-MCI (82)</td>
<td>P-MCI (75)</td>
<td>P-MCI12 (28)</td>
<td>P-MCI12 (47)</td>
</tr>
<tr>
<td>r</td>
<td>3.55 ± 3.02</td>
<td>5.33 ± 3.29</td>
<td>5.33 ± 3.45</td>
</tr>
<tr>
<td>l</td>
<td>3.46 ± 3.30</td>
<td>6.08 ± 3.18</td>
<td>6.43 ± 3.72</td>
</tr>
<tr>
<td>r+l</td>
<td>3.50 ± 2.90</td>
<td>5.70 ± 2.96</td>
<td>5.86 ± 3.36</td>
</tr>
</tbody>
</table>
Correlation with clinical values

We used Set 1 (555 subjects with Month 12 follow-up) to determine correlations of atrophy with clinical data. Table 5 shows Pearson’s r-values for the correlation of atrophy with MMSE, CDR, and the change of both values over one year. Correlations are displayed for the whole image set as well as for the clinical groups individually.

Since CDR does not vary within the MCI and CN groups at Baseline, no meaningful correlation can be measured. When using all subjects, a significant correlation in the anticipated direction could be observed in all tests. Correlations were almost as strong for the MCI group and were still significant for the left hippocampus when looking at the AD group separately.

ApoE genotype

Further tests were carried out to gain an understanding of the influence of a subject’s ApoE genotype (determined by the ApoE alleles carried) on hippocampal atrophy. Humans carry two out of three possible ApoE alleles ($\varepsilon_2$, $\varepsilon_3$, and $\varepsilon_4$). Carriers of ApoE4 have been shown to have a higher risk of developing AD, while ApoE2 carriers have a lower risk (Lehtovirta et al., 1996). Table 6 shows the results of a two-tailed t-test comparing the atrophy rates for $\varepsilon_3/3$ and $\varepsilon_3/4$ carriers ($\varepsilon_2/2$, $\varepsilon_2/3$, $\varepsilon_2/4$, and $\varepsilon_4/4$ carriers were excluded). Significant differences between the genotypes can be observed when looking at all subjects simultaneously, but also within subgroups—controls, MCI and the combination of both. No significant difference of atrophy rates in the left hippocampus can be observed when only looking at the control group.

Additionally, we compared atrophy rates for the $\varepsilon_2/3$ and $\varepsilon_3/3$ carriers. The only fairly significant difference in atrophy rates, however, could be observed for the left hippocampus when using all subjects ($t = 2.28, p = 0.02$) or when combining CN and MCI groups ($t = 2.13, p = 0.03$).

Discrimination between clinical groups based on atrophy

We evaluated the automatically determined atrophy values for their power to discriminate between subject groups. Receiver operating characteristic (ROC) curves for atrophy-based classification after 12 and 24 months are displayed in Fig. 4.

Table 5

<table>
<thead>
<tr>
<th></th>
<th>All (555)</th>
<th>CN (163)</th>
<th>MCI (268)</th>
<th>AD (124)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMSE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>r</td>
<td>0.43*</td>
<td>−0.13</td>
<td>−0.31*</td>
<td>−0.17</td>
</tr>
<tr>
<td>l</td>
<td>−0.52*</td>
<td>0.09</td>
<td>−0.38*</td>
<td>−0.26*</td>
</tr>
<tr>
<td>ΔMMSE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>r</td>
<td>0.30*</td>
<td>0.16</td>
<td>0.26*</td>
<td>0.13</td>
</tr>
<tr>
<td>l</td>
<td>0.36*</td>
<td>0.14</td>
<td>0.32*</td>
<td>0.22b</td>
</tr>
<tr>
<td>CDR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>r</td>
<td>0.38*</td>
<td>N.A.</td>
<td>N.A.</td>
<td>0.14</td>
</tr>
<tr>
<td>l</td>
<td>0.47*</td>
<td>N.A.</td>
<td>N.A.</td>
<td>0.22b</td>
</tr>
<tr>
<td>ΔCDR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>r</td>
<td>−0.21*</td>
<td>−0.06</td>
<td>−0.15b</td>
<td>−0.15</td>
</tr>
<tr>
<td>l</td>
<td>−0.27*</td>
<td>−0.08</td>
<td>−0.20a</td>
<td>−0.23b</td>
</tr>
</tbody>
</table>

Table 6

<table>
<thead>
<tr>
<th></th>
<th>All (96/42)</th>
<th>MCI (115/141)</th>
<th>CN and MCI (211/183)</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>−6.09*</td>
<td>−3.21*</td>
<td>−2.95*</td>
</tr>
<tr>
<td>l</td>
<td>−5.33*</td>
<td>−1.1</td>
<td>−2.60*</td>
</tr>
</tbody>
</table>

Fig. 2. Hippocampal volume loss in % from Baseline after 12 and 24 months. Box-and-whisker plots for AD, P-MCI, S-MCI, CN.

Fig. 3. Comparison of volume loss after 12 months when segmenting two (method a) or three (method b) timepoints simultaneously. Dashed lines represent the 95% confidence interval of the mean (solid line).
A bootstrapping approach that has previously been used for classification based on hippocampal volume (Chupin et al., 2009) was used to evaluate the classification rate between pairs of clinical groups: for each group 75% of the subjects were randomly selected for training. The remaining 25% were then classified according to their difference from the mean rates estimated in the training sets. The average classification rate, sensitivity and specificity for different groups after 5000 runs is displayed in Table 7. Values based on atrophy rates after 24 months are given in parentheses. Using atrophy rates from the first year of observation, a classification rate of 75%–82% is obtained when discriminating between healthy controls and AD patients or subjects that develop AD during the study. Of clinical interest is the identification of subjects converting from MCI to AD. Early and reliable detection of these subjects could support clinical decisions for or against therapy with disease-modifying drugs. Hippocampal atrophy over the course of the study. Of clinical interest is the identification rate of atrophy. With a chosen effect size of 0.88 (0.92), 0.71 (0.77), 0.83 (0.86), and 0.72 (0.71), respectively. AUC’s for rates after 24 months are given in parentheses.

Using atrophy rates from the first year of observation, a classification rate of 75%–82% is obtained when discriminating between healthy controls and AD patients or subjects that develop AD during the study. Of clinical interest is the identification of subjects converting from MCI to AD. Early and reliable detection of these subjects could support clinical decisions for or against therapy with disease-modifying drugs. Hippocampal atrophy over the first year correctly identified 70% of subjects who converted from MCI to AD in the same period. An even more interesting result is the classification rate of 64% between subjects who did not convert within the entire observation period and subjects who converted after 12 months. Taking atrophy rates after 2 years, better results are achieved in all pairings except P-MCI ≤ 12 vs S-MCI.

Sample size calculation

For each patient group, we estimated the sample size needed in a hypothetical two-arm study to detect a reduction in the mean annual rate of atrophy. With a chosen effect size of Δ and a standard deviation σ, the following formula can be used to estimate the sample size needed:

\[ n = \frac{2\alpha^2 (z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2}. \]  

We chose the Δ to be 0.25 μ where μ is the mean atrophy rate of the corresponding group (see Tables 3 and 4). We set the significance level (α) to 0.05 and the power (1 – β) to 0.8. The cutoff points of the standard normal probability distributions matching the defined significance and statistical power are \( z_{1-\alpha/2} \approx 1.96 \) and \( z_{1-\beta} \approx 0.84 \) respectively.

The total estimated sample sizes for both arms needed to detect a 25% reduction in the AD and MCI groups in intervals of 12 and 24 months are displayed in Table 8.

Segmentation accuracy

Test re-test reliability

To test the reliability of the proposed method, we applied single timepoint segmentation to ten image pairs that were each acquired from the same ADNI subject in the same study session. When randomly selecting a reference segmentation for each pairing, the average volume difference to the second segmentation is not statistically significant. The average absolute difference between the measurements is 1.2 ± 1.3% of their average value. Applying 4D graph cuts to these image pairs reduces the average absolute difference to 0.34 ± 0.36%. This shows that segmentations obtained simultaneously from multiple time points are more consistent than single-time point segmentations.

Comparison of simultaneous to semi-automatic independent segmentation

To assess the importance of segmenting images from all time-points simultaneously, we compared our atrophy estimates with those based on the label maps provided by ADNI as described in the Hippocampus atlases section. These label maps have previously been used to study hippocampal atrophy in work by Schuff et al. (2009).

<table>
<thead>
<tr>
<th>AD/CN</th>
<th>MCI/CN</th>
<th>P-MCI/CN</th>
<th>P-MCI ≤ 12/CN</th>
<th>P-MCI ≤ 12/S-MCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class rate</td>
<td>82% (86%)</td>
<td>63% (72%)</td>
<td>76% (83%)</td>
<td>80% (82%)</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>81% (85%)</td>
<td>59% (65%)</td>
<td>73% (79%)</td>
<td>76% (69%)</td>
</tr>
<tr>
<td>Specificity</td>
<td>83% (87%)</td>
<td>71% (83%)</td>
<td>78% (85%)</td>
<td>81% (86%)</td>
</tr>
<tr>
<td>Normal vs AD</td>
<td>Normal vs MCI</td>
<td>Normal vs P-MCI</td>
<td>P-MCI vs S-MCI</td>
<td></td>
</tr>
<tr>
<td>AD/CN</td>
<td>MCI/CN</td>
<td>P-MCI/CN</td>
<td>P-MCI ≤ 12/CN</td>
<td>P-MCI ≤ 12/S-MCI</td>
</tr>
<tr>
<td>Class rate</td>
<td>66% (67%)</td>
<td>70% (67%)</td>
<td>64% (68%)</td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>62% (66%)</td>
<td>57% (61%)</td>
<td>63% (70%)</td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td>68% (69%)</td>
<td>72% (70%)</td>
<td>64% (68%)</td>
<td></td>
</tr>
</tbody>
</table>

The hypothesis that the distribution has zero mean cannot be rejected with p = 0.32.
We used the similarity index (SI) (Dice, 1945) to measure average overlaps between the manually corrected label maps and the segmentation produced by the proposed method. The average overlap for 262 Baseline and Month 12 follow-up images is 0.83 ± 0.04. There is no significant difference between left and right hippocampus.

We used the subset of images for which label maps at Baseline and Month 12 were provided by ADNI to compare both approaches of atrophy measurement. Resulting atrophy rates (%) are shown in Table 9. Despite the significant differences in mean values, there is a high correlation between both measures with ρ = 0.45 and p < 0.001 when looking at all values. The correlation is still high and significant (ρ = 0.001) when looking at AD and MCI groups separately (ρ = 0.61, r = 0.42 respectively).

Fig. 5 shows ROC curves, demonstrating the ability of both measurements to discriminate between clinical groups. Although the mean difference between clinical groups is higher with the ADNI label maps, classification results are better with the proposed graph cuts approach. This can be explained by the larger precision of the proposed method, evidenced by the lower standard deviation of the atrophy rates measured. The sample sizes required to detect a 25% reduction in atrophy in the AD and MCI groups confirm this observation with substantially lower values for the proposed method. Atrophy rates based on the label maps provided by ADNI result in samples sizes for both arms of 150 and 750 subjects for the AD and MCI groups respectively. Applying 4D graph cuts to this subset results in reduced sample sizes of 62 and 204 subjects respectively.

Temporal smoothness term
To assess the influence of the weighting factor αpq introduced in the Energy terms section that weights spatial and temporal edges individually, we run the atrophy measurement over 12 months for Set 1 with different parameter settings. While small parameter changes do not influence the segmentation outcome substantially, we found that weighting spatial edges with around 20 times higher than temporal edges leads to a robust framework that results in consistent segmentations but is still flexible enough to accurately detect atrophy. Depending on the structure to be segmented and expected difference over time, temporal constraints can be varied in different settings.

Setting αpq = 0 for all temporal edges leads to average atrophy rates of 3.94 ± 2.13, 2.32 ± 2.31, 0.87 ± 1.66 for the AD, MCI and CN groups respectively. The increased difference in mean values does not outweigh the increase in standard deviation and therefore results in slightly worse classification results and larger sample sizes needed to detect change.3

Discussion and conclusion
We applied a 4D graph cuts segmentation method to measuring hippocampal atrophy in longitudinal MR images from AD patients, subjects with MCI as well as age matched healthy controls taken from the ADNI study. We simultaneously segmented 568 image pairs (Baseline and Month 12 follow-up) as well as 362 image triplets (Baseline, Month 12, Month 24 follow-up). The resulting atrophy rates confirm previous results for hippocampal loss in AD and healthy aging, with atrophy rates significantly higher in AD (3.85 ± 1.99 vs. 0.85 ± 1.59). The values are in the same range as atrophy rates for both groups reported in a recently published meta-analysis of hippocampal loss rates in AD which combines nine studies using manual and automatic approaches (Barnes et al., 2009). Two recent studies report substantially different atrophy rates for a similar subset of ADNI subjects: Morra et al. (2009) with AD: 5.59 ± 7.24, CN: 0.66 ± 5.96 and Schuff et al. (2009) with AD: 4.4 ± 5.88, CN: 0.8 ± 5.63.4 While the hippocampus atlases we used are based on the same protocol used in Schuff et al. (2009), the differences to the atrophy rates reported in Morra et al. (2009) may partly be explained by a potential difference in region definition. In addition, both previous studies report relatively large confidence intervals which make an estimate of mean values less reliable.

We found that atrophy rates in subjects with progressive MCI are significantly higher than in subjects with a stable diagnosis of MCI. Furthermore, subjects with stable MCI show higher atrophy rates than control subjects. These results confirm findings by Wang et al. (2009) and are also supported by the finding of significantly reduced cortical thickness in the P-MCI group compared to the S-MCI group reported in Julkunen et al. (2009). Our results furthermore show that subjects converting to AD during the first year of the study showed significantly higher atrophy in that time period. More interesting, however, is the significantly higher atrophy rate of subjects converting to AD after year one. This suggests that substantial loss in hippocampal volume can be observed before a conversion to AD is diagnosed with psychological tests.

We used the automatically determined atrophy rates over 12 months to determine their correlation with clinical variables, comparing our results to previously reported values using a similar subset of ADNI images (Morra et al., 2009). A direct numerical comparison of both methods is not possible. Stronger and statistically more significant correlations indicate, however, that the method we propose achieves better accuracy. When using all subjects, a strong and highly significant correlation between atrophy rates and MMSE, CDR as well as the change of both variables over time could be observed. Taking into account the definition of these clinical variables and the difference in atrophy reported, these correlations are as expected. When looking at the MCI group separately, the correlation is almost as significant. In the AD group, however, only a relatively poor correlation between atrophy rates and clinical variables was measured for the left hippocampus. This confirms findings by Morra et al. (2008). Apart from the lower power to detect correlation caused by the relatively small group size, a potential explanation is the heterogeneity of the AD group with respect to change in clinical variables (see Table 1). The absence of a significant correlation for the control group can probably be explained by the small amount of variation of both atrophy rates and clinical variables.

---

3 Classification was performed as described in Discrimination between clinical groups based on the atrophy section, results are not shown here. Using Eq. (4) shows slightly higher sample sizes compared to the ones reported in the Sample size calculation section.

4 Standard deviations were calculated from 95% confidence intervals and standard errors respectively as well as sample sizes provided in the original work.

### Table 8
Estimated sample sizes for both arms that would be needed to detect a 25% reduction in atrophy in the AD and MCI groups in intervals of 12 and 24 months.

<table>
<thead>
<tr>
<th>Interval</th>
<th>AD</th>
<th>MCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 months</td>
<td>67</td>
<td>206</td>
</tr>
<tr>
<td>24 months</td>
<td>46</td>
<td>121</td>
</tr>
</tbody>
</table>

### Table 9
Average atrophy rates (%) for the subset of Set 1 for which hippocampal label maps were provided by ADNI. Atrophy rates based on these label maps are compared to automatically determined rates based on the proposed method. Numbers of subjects are given in parentheses. mean±std.

<table>
<thead>
<tr>
<th></th>
<th>CN (85)</th>
<th>MCI (122)</th>
<th>S-MCI (65)</th>
<th>P-MCI (57)</th>
<th>AD (55)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADNI labels</td>
<td>1.10 ± 5.82</td>
<td>3.23 ± 5.58</td>
<td>2.72 ± 5.49</td>
<td>3.81 ± 5.66</td>
<td>6.27 ± 4.84</td>
</tr>
<tr>
<td>4D graph cuts</td>
<td>0.9 ± 1.61</td>
<td>2.31 ± 2.08</td>
<td>1.82 ± 1.89</td>
<td>2.87 ± 2.16</td>
<td>3.67 ± 1.82</td>
</tr>
</tbody>
</table>
Furthermore, we investigated the influence of a subject's ApoE gene status on hippocampal atrophy. We found a statistically highly significant difference between ε3/3 and ε3/4 carriers when combining all subject groups. Remarkably, the difference is still significant when looking at control or MCI groups only. We also looked at the difference in atrophy rates of ε2/3 and ε3/3 carriers, but could only observe trends.

The reported atrophy rates allow classification between subject groups. Although to our knowledge no classification results based on hippocampal atrophy have been published for the ADNI group so far, other classifiers have been proposed. Based on baseline volume of the hippocampus, Chupin et al. (2009) report a rate of 64% for the clinically important classification between MCI converters (P-MCI) and subjects with stable MCI (S-MCI). In Gerardin et al. (2009), a more sophisticated classifier based on hippocampal shape features achieves discrimination between MCI and controls with an accuracy of 80%.

Hippocampal atrophy rates over 12 months based on 4D graph cuts distinguish between controls and AD or MCI with a classification rate of 82% and 67% respectively. A discrimination of MCI converters from healthy subjects and especially from MCI non-converters is of clinical importance. With the proposed method, all converters could be discriminated from controls with a rate of at least 75%. Atrophy rates after 12 months allow the identification of 70% of the subjects that convert from MCI to AD in the same period. The classification rate of 64% between non-converters and subjects that converted after Month 12 shows that an indication of future conversion can be obtained before clinical tests identify the subjects as AD patients. Taking atrophy rates after 2 years, better results are achieved in all pairings except P-MCI vs S-MCI. This can probably be partly explained by missing information about subjects that progress from the S-MCI group to the P-MCI group after 24 months. Although all subjects are followed for 36 months in the ADNI study, the final examination is not available for the majority of subjects. Some subjects are likely to convert to AD after Month 24 but are assigned to the S-MCI group, which spuriously reduces the classification rates. Another factor is probably the relatively small sample size for the interval between Month 12 and 24 (especially for P-MCI vs S-MCI), which results in relatively large confidence intervals around the mean atrophy rate (see Table 4).

We found a high level of agreement between the individual hippocampal segmentations generated by the proposed method and semiautomatically generated reference segmentations provided by ADNI (SI 0.83 ± 0.04). Atrophy rates calculated on the basis of both methods were strongly correlated. Significant differences between the two approaches are seen when the comparison is based on classification rates and statistical power: on these criteria, the 4D graph cuts based method is clearly superior. We attribute this superiority to the presence of increased temporal constraints when segmenting images of all timepoints simultaneously: this leads to higher consistency within the ensembles of measurements on which the atrophy classification is based.

In future studies we plan to apply the proposed method to atrophy measurement in other brain structures than the hippocampus by using a segmentation based on a more detailed anatomical atlas, e.g. Hammers et al. (2003). Atrophy rates of different regions could then be used to form a potentially stronger classifier for an early detection of Alzheimer’s disease.

Acknowledgments

This work was partially funded under the 7th Framework Programme by the European Commission (http://cordis.europa.eu/ist; EU-Grant-224328-PredictAD; Name: From Patient Data to Personalised Healthcare in Alzheimer’s Disease).

RAH was supported by a research grant from the Dunhill Medical Trust.

Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI; Principal Investigator: Michael Weiner; NIH grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering (NIBIB), and through generous contributions from the following: Pfizer Inc., Wyeth Research, Bristol-Myers Squibb, Eli Lilly and Company, GlaxoSmithKline, Merck & Co. Inc., AstraZeneca AB, Novartis Pharmaceuticals Corporation, Alzheimer’s Association, Eisai Global Clinical Development, Elan Corporation plc, Forest Laboratories, and the Institute for the Study of Aging, with participation from the U.S. Food and Drug Administration. Industry partnerships are coordinated through the Foundation for the National Institutes of Health. The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory of Neuroimaging at the University of California, Los Angeles.

Appendix A. The Alzheimer’s Disease Neuroimaging Initiative

The Alzheimer’s Disease Neuroimaging Initiative (ADNI) was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB),
the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public–private partnership. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principle Investigator of this initiative is Michael W. Weiner, M.D., VA Medical Center and University of California — San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55 to 90, to participate in the research — approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years, and 200 people with early AD to be followed for 2 years. For up-to-date information see www.adni-info.org.

Image acquisition and preprocessing

Image acquisition was carried out at multiple sites based on a standardized MRI protocol (Jack et al., 2008) using 1.5 T scanners manufactured by General Electric Healthcare (GE), Siemens Medical Solutions, and Philips Medical Systems. Out of two available 1.5 T T1-weighted IR images based on a 3D MPRAGE sequence, we used the image that has been designated as “best” by the ADNI quality assurance team (Jack et al., 2008). Acquisition parameters on the Siemens scanner (parameters for other manufacturers differ slightly) are echo time (TE) of 3.924 ms, repetition time (TR) of 8.916 ms, inversion time (TI) of 1000 ms, flip angle 8°, to obtain 166 slices of 1.2-mm thickness with a 256 × 256 matrix.

All images were preprocessed by the ADNI consortium using a pipeline consisting of GradWarp (A system-specific correction of image geometry distortion due to gradient non-linearity (Jovicich et al., 2006)), B1 non-uniformity correction (Correction for image intensity non-uniformity (Jack et al., 2008)) and N3 (A histogram peak sharpening algorithm for bias field correction (Sled et al., 1998)). Since the Philips systems used in the study were equipped with B1 correction and their gradient systems tend to be linear (Jack et al., 2008), the first two preprocessing steps were applied by ADNI only to images acquired on GE and Siemens scanners.

References


