Pathology-related Automated Hippocampus Segmentation Accuracy

M. Liedlgruber[2], K. Butz[1,4], Y. Höller[1], G. Kuchukhidze[1], A. Taylor[1], A. Thomschewski[1,4], O. Tomasi[3], E. Trinka[1,4] und A. Uhl[2]

[1] Department of Neurology, Christian Doppler Medical Centre and Centre for Cognitive Neuroscience, Paracelsus Medical University, Salzburg, Austria
[2] Department of Computer Sciences, University of Salzburg, Austria
[3] Department of Neurosurgery, Paracelsus Medical University, Salzburg, Austria
[4] Spinal Cord Injury and Tissue Regeneration Center Salzburg, Austria
uhl@cosy.sbg.ac.at

Abstract Hippocampal segmentation accuracy of out-of-the-box software tools (FreeSurfer, AHEAD, BrainParser) is analysed wrt. potential variability in populations with different pathologies. Findings confirm variabilities wrt. different pathologies but also human rater ground truth and single pathologies exhibit significant variability as well.

1 Introduction

Mild cognitive impairment (MCI) is a condition of cognitive deterioration that is difficult to classify as normal aging or as a prodromal stage to dementia [1]. Despite considerable progress of research, current endeavours are still focused on accurate and early diagnosis of MCI [2]. Similarly, the diagnosis of temporal lobe epilepsy (TLE) was [3] and still is based on clinical assessment and electroencephalographic (EEG) examination, sometimes being inconclusive. Both diseases need to be treated and handled adequately, in order to prevent massive memory decline or hazards by seizures. From a structural point of view, the hippocampus is an area of the brain that links the two conditions [4]. It is therefore worth evaluating techniques for the diagnosis of these conditions that are based on distinctive features of this structure of the brain. Segmentation of the hippocampi is of course a prerequisite for such approaches.

Since manual definition of the borders of the hippocampus is tedious and time-consuming work, many techniques for automated hippocampus segmentation have been published over the last years [5,6,7] including state-of-the-art algorithms based on multi-atlas segmentation (MAS)[8].

In this paper, we look into the accuracy of out-of-the-box segmentation software (which is highly attractive for research groups interested in segmentation results but not segmentation algorithm development) applied to hippocampi when differentiating patients diagnosed with MCI and TLE and comparing the results to a healthy control group. Previous work [9] revealed significant variability wrt. various aspects of hippocampal segmentation however being restricted to an overall analysis without differentiating different subject groups. As a hypothesis, one might conjecture that automated segmentation tools tend to commit more errors in subjects suffering from MCI or TLE as the atlases they base their segmentation on usually consist of healthy subjects. In Section 2, we describe the employed data set (including the rare availability of a three human rater ground truth significantly surpassing [9]) and automated segmentation tools used in this study. Section 3 outlines experimental setup and present results, while a conclusion is given in Section 4.

2 Materials and Methods

2.1 Dataset and Manual Ground Truth

In this work we use a data set that has been acquired at the Department of Neurology, Paracelsus Medical University Salzburg and consists of 58 T1-weighted MRI volumes, including patients with mild cognitive impairment (MCI, 20 subjects), with temporal lobe epilepsy (TLE, 17 subjects), and a healthy control group (CG, 21 subjects). The dataset contains 28 males (18-76 years, mean age 53±19 years) and 30 females (23-71 years, mean age 54±14 years). We defined patients with amnestic MCI according to level three of the global deterioration scale for aging and dementia. Diagnosis was based on multimodal neurological assessment, including imaging (high resolution 3T magnetic resonance tomography, and single photon emission computed tomography with Hexamethylpropylenaminooxim), and neuropsychological testing.

Manual segmentations have been performed by 3 experienced raters (one senior neurosurgeon and two junior neuroscientists supervised by a senior neuroradiologist) on a Wacom Cintiq 22HD graphic tablet device (resolution 1920x1200) using a DTK-2200 pen and employing the 32-bit 3DSlicer software for Windows (v. 4.2.2-1 r21513) to delineate hippocampus voxels for each slice separately. The raters independently used consensus on anatomical landmarks/boarders of the hippocampus based on Henry Duvemoy's hippocampal anatomy [10]. The procedure used was to depict the hippocampal outline in the view of all planes in the following order: sagittal − coronal − axial with subsequent cross line control through all planes.

2.2 Hippocampus Segmentation Software Packages

In contrast to most of the algorithms presented in literature, e.g. [7], all employed software packages are already pre-compiled and available for free [9]:
FreeSurfer (FS)[1] is a set of tools which allow an automated labelling of subcortical structures in the brain. Such a subcortical labelling is obtained by using the volume-based stream which consists of five stages [5]. The result is a label volume, containing labels for various different subcortical structures (e.g. hippocampus, amygdala, and cerebellum). FreeSurfer is a highly popular tool to

[1] v. 51.0, available at http://surfer.nmr.mgh.harvard.edu

assess clinical hypotheses [11] or to compare to newly proposed segmentation techniques (e.g. [7]).
AHEAD (Automatic Hippocampal Estimator using Atlas-based Delineation[2]) is specifically targeted at an automated segmentation of hippocampi [6]. Based on multiple atlases and a statistical learning method, the final segmentation is obtained.
Although **BrainParser (BP)**[3] is usually able to label various different subcortical structures, we use a version which is specifically tailored to hippocampus segmentation. The tool uses a deformable registration between the input and the reference volume and subsequent corresponding input volume labeling.

In case of BrainParser and AHEAD the MNI152 atlas has been used as provided with the software. For FreeSurfer we used the MNI305 atlas.

2.3 Metrics Used to Assess Segmentation Quality

In the following the automated segmentation is denoted by S, the ground truth segmentation is called G, and $v(\cdot)$ is a volume operator which computes the volume of a voxel volume with respect to the actual dimensions of a voxel.

- **Symmetric Hausdorff distance (SHD)**
  This metric is based on the actual structure of a voxel volume. It is defined as
  $$SHD(G,S) = \max(HD(G,S), HD(S,G)) \tag{1}$$
  where
  $$HD(X,Y) = \max_{x \in X}(\min_{y \in Y} d(x,y)). \tag{2}$$
  is the non-symmetric Hausdorff distance, x and y are vectors in $\mathbb{R}^3$ and $d(\cdot,\cdot)$ denotes the Euclidean distance between two vectors.
- **Relative overlap (RO)**
  The relative overlap (also known as the Jaccard similarity coefficient) represents the fraction of voxels in the union of G and S which are also contained in the intersection of G and S.
  $$RelativeOverlap(G,S) = \frac{v(G \cap S)}{v(G \cup S)} \tag{3}$$

While low values in $[0,1]$ correspond to little similarity / quality for RO, the SHD produces large values (differences) between dissimilar segmentations.

3 Results

The following results are always based on both hippocampi simultaneously (both hippocampi from each scan are treated as one segmentation object).

[2] version 1.0, available at http://www.nitrc.org/projects/ahead
[3] available at http://www.nitrc.org/projects/brainparser

First, we provide quantitative results in terms of normalised hippocampus volumina in Table 1 (i.e. the percentage of the entire brain volume of each subject is given). As the hippocampus in known to be atrophic in MCI and dementia [12] and is sclerotic in specific subtypes of epilepsy [13] we expect reduced volumina for MCI and TLE (V_{MCI}, V_{TLE}) respectively, as compared to the volume of healthy subjects (V_{CG}).

We clearly have $V_{CG} > V_{TLE} > V_{MCI}$ as the main result seen for all three raters consistently, thus, results are corresponding well with the expectations at first sight. However, this is only true when looking at the raters' results individually. However, cross-rater differences are significant and volumes among raters vary by up to 20%. Additionally, partially high standard deviations among subjects obliterate the clear trend as seen from the averaged values.

Table 1. Summary of normalised hippocampus volumina (in percent of the entire brain volume) as obtained from the three human raters and the three software packages, averaged over all subjects of each pathology class also showing standard deviation.

	V_{CG}	V_{MCI}	V_{TLE}
Rater 1	0.482±0.092	0.383±0.065	0.442±0.060
Rater 2	0.450±0.125	0.360±0.079	0.393±0.078
Rater 3	0.556±0.086	0.457±0.058	0.514±0.073
FreeSurfer	0.542±0.135	0.499±0.086	0.579±0.120
AHEAD	0.333±0.031	0.304±0.049	0.348±0.075
BrainParser	0.366±0.116	0.363±0.057	0.376±0.067

For the automated segmentation tools we observe a different, still clear ordering as displayed in the table: $V_{TLE} > V_{CG} > V_{MCI}$. This of course does not correspond to the expectations. While the known over-segmentation of FS [11] is also reflected in our results (making the comparison with AHEAD and BP volumina impossible), we also find clear cross-tool variation between AHEAD and BP as well as significant standard deviations (see e.g. FS and BP for CG and FS for TLE). Thus, in terms of volumina, human rater results are closer to the expected values, however, for both human raters as well as automated segmentation tools we notice significant inter-rater and inter-tool variability as well as high subject variability as indicated by high standard deviations.

The following Table 2 provides a more qualitative view when comparing the segmentation results. We compare the results of the automated tools with the human rater ground truth (which is a voxel-based majority vote among the three raters) in terms of the two metrics, SHD and RO, respectively. Apart from differences in volume also shape differences are reflected by these metrics, where SHD indicates shape differences in the most pronounced manner.

The upper half of Table 2 shows results wrt. SHD. FS results meet the expectations in that lowest distance to human raters is seen for CG subjects. The largest distance (i.e. error) to the ground truth is seen for TLE patients. The other two segmentation tools exhibit a different behaviour: While the relation between TLE and MCI patients is identical to FS segmentations, CG subjects exhibit the largest distance to the ground truth. While this result is highly unexpected, we need to consider the extremely high standard deviation in the CG results of AHEAD and BP. It seems that the data set contains CG subjects for which those two segmentation tools are highly erroneous, while for others the results are quite good.

Table 2. Summary of segmentation assessment metrics (SHD and RO) computed between the automated segmentations and the ground truth (majority voted among three raters), averaged over all subjects of each pathology class.

	Overall	**CG**	**MCI**	**TLE**
SHD Results				
FreeSurfer	7.73±2.27	6.95±1.36	7.74±1.98	8.93±3.00
AHEAD	7.78±14.03	11.31±24.13	5.30±0.98	6.89±2.61
BrainParser	9.64±16.29	14.43±26.87	7.38±9.11	8.10±3.41
RO Results				
FreeSurfer	0.63±0.06	0.66±0.03	0.62±0.05	0.59±0.09
AHEAD	0.62±0.08	0.59±0.12	0.64±0.03	0.61±0.05
BrainParser	0.59±0.17	0.54±0.23	0.61±0.15	0.59±0.09

The lower half of Table 2, showing the results of the RO metric, basically confirms the findings of the SHD metric. Again, FS results corresponds to the expectations (higher similarity for CG subjects as compared to MCI and TLE patients), while AHEAD and BP results show the CG subjects as those with lowest similarity to the ground truth. By perfect analogy to the SHD metric also RO results rate segmentations of MCI patients more similar to ground truth as compared to TLE patients **and** we observe very high standard deviations in the metric values for AHEAD and BP considering CG subjects.

4 Discussion

For FS segmentations, we find more errors in subjects suffering from MCI and TLE, compared to errors in the CG population. However, AHEAD and BP segmentations exhibit lowest correspondence to human ground truth for CG subjects on average, although for this population a very high standard deviation is present in the results. All three tools commit more errors in subjects suffering from TLE as compared to the MCI patient population.

All these results have to be taken with great caution, as the variability in the results of the three human raters is found to be very high, i.e. the inter-rater variability in terms of hippocampal volume of the same pathology class (CG, MCI, or TLE) is in the same order of magnitude as the inter-pathology volume difference of a single rater.

Acknowledgments

This work has been funded by the Austrian Science Fund (FWF) under Project No. KLI 00012.

References

1. Hänninen T, Soininen H. Age-associated memory impairment. Normal aging or warning of dementia? Drugs Aging. 1997;11:480–9.
2. Lei B, Chen S, Ni D, Wang T. Discriminative Learning for Alzheimer's Disease Diagnosis via Canonical Correlation Analysis and Multimodal Fusion. Frontiers in Aging Neuroscience. 2016;8(77).
3. Tharp B. Recent progress in epilepsy - Diagnostic procedures and treatment. Calif Med. 1973;119:19–48.
4. Höller Y, Trinka E. What do temporal lobe epilepsy and progressive mild cognitive impairment have in common? Front Syst Neurosci. 2014;8(58).
5. Fischl B, van der Kouwe A, Destrieux C, et al. Automatically Parcellating the Human Cerebral Cortex. Cerebral Cortex. 2004;14(1):11–22.
6. Suh JW, Wang H, Das S, Avants B, Yushkevich PA. Automatic Segmentation of the Hippocampus in T1-Weighted MRI with Multi-Atlas Label Fusion Using Open Source Software: Evaluation in 1.5 and 3.0T ADNI MRI. In: Proceedings of the International Society for Magnetic Resonance in Medicine conference (ISMRM'11); 2011. p. 3844.
7. Zarpalas D, Gkontra P, Daras P, Maglaveras N. Accurate and Fully Automatic Hippocampus Segmentation Using Subject-Specific 3D Optimal Local Maps Into a Hybrid Active Contour Model. IEEE Journal of Translational Engineering in Health and Medicine. 2014;2:1–16.
8. Leung KK, Barnes J, et al. Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. NeuroImage. 2013;51(4):1345–1359.
9. Liedlgruber M, Butz K, Höller Y, Kuchukhidze G, Taylor A, Tomasi O, et al. Variability Issues in Automated Hippocampal Segmentation: A Study on Out-of-the-Box Software and Multi-rater Ground Truth. In: Proceedings of the 29th IEEE International Symposium on Computer-Based Medical Systems (CBMS'16); 2016. p. 191–196.
10. Kuzniecky R, Jackson GD. Magnetic resonance in epilepsy. New York: Raven Press; 1995.
11. Cherbuin N, Anstey1 KJ, Réglade-Meslin C, Sachdev PS. In Vivo Hippocampal Measurement and Memory: A Comparison of Manual Tracing and Automated Segmentation in a Large Community-Based Sample. PLoS ONE. 2009;4:1–10.
12. Fotuhi M, Do D, Jack C. Modifiable factors that alter the size of the hippocampus with ageing. Nat Rev Neurol. 2012;8:189–202.
13. Malmgren K, Thom M. Hippocampal sclerosis - origins and imaging. Epilepsia. 2012;53:19–33.