Assessing Out-of-the-box Software for Automated Hippocampus Segmentation*

M. Gschwandtner[2], Y. Höller[1], M. Liedlgruber[2], E. Trinka[1] 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
uhl@cosy.sbg.ac.at

Abstract A comparison of four out-of-the-box software packages for automated hippocampus segmentation reveals that AHEAD and Freesurfer deliver the most satisfying results in terms of software usability and segmentation reliability and are thus recommended to be used in a fused manner.

1 Introduction

The hippocampus is reduced in size in individuals with obesity, diabetes mellitus, hypertension, hypoxic brain injury, obstructive sleep apnoea, bipolar disorder, clinical depression, and head trauma, it is atrophic in mild cognitive impairment and dementia [1], and it is sclerotic in specific subtypes of epilepsy [2]. The most established application of hippocampus volumetry is the prediction of conversion from normal aging to mild cognitive impairment, and further to Alzheimer disease [3].

Since manual definition of the borders of the hippocampus is tedious and time-consuming work, many techniques and algorithms for automated hippocampus segmentation have been published over the last years [4,5,6]. However, without proper background and significant experience, a re-implementation of these techniques is far from being trivial and usually requires several man-years of programming effort. Therefore, especially for research groups "only" interested in segmentation results for further analysis, available (preferably cost-free) out-of-the-box segmentation software without the need for extensive optimisation and adaption is a highly attractive (if not the only) option.

In this paper, we assess four cost-free and pre-compiled out-of-the-box hippocampus segmentation software packages, both from a usability aspect (installation effort and computational cost) as well as from a segmentation accuracy viewpoint. In Section 2, the employed test dataset is described and properties of the four software packages considered are reviewed. Section 3 outlines experimental setup and present results, while a conclusion and recommendation is given in Section 4.

* This work has been supported by the Austrian Science Fund under Project No. KLI 00012.

2 Methodology and Software

2.1 Dataset

The database of MRI scans used is the Radiology Research Database[1] [7], which consists of 50 brain T1-weighted MRI volumes. Forty of these volumes belong to patients with temporal lobe epilepsy (TLE), potentially having atrophic hippocampi. The remaining ten subjects are non-epileptic. Since the ground truth segmentation is provided only for the first 25 subjects in this database (all TLE patients), we restrict our evaluation to these subjects (from now on we refer to this subset when referring to the Radiology Research Database). The subset consists of 7 males (27-55 years, mean age 40±11 years) and 18 females (15-55 years, mean age 36±12 years).

2.2 Software Packages Evaluated

We evaluated four different software packages in the context of an automated segmentation of hippocampi. In contrast to most of the algorithms presented in [6], all these software packages are already pre-compiled and available for free.

FreeSurfer[2] is a set of tools which allow an automated labeling of subcortical structures in the brain. Such a subcortical labeling is obtained by using the volume-based stream which consists of five stages [4]. After an affine registration of the volume with Talairach space and a bias field correction, an initial labeling of the voxels is obtained (based on the atlas template priors). This is followed by a non-linear alignment of the volume to the atlas. To obtain the final segmentation, the initial segmentation is refined iteratively by computing probabilities (based on the label of a voxel and the labels of the neighboring voxels) and performing a re-segmentation. The result is a label volume, containing labels for various different subcortical structures (e.g. hippocampus, amygdala, and cerebellum).

AHEAD (Automatic Hippocampal Estimator using Atlas-based Delineation[3]) is specifically targeted at an automated segmentation of hippocampi [5]. After an initial rigid registration step, a deformable registration is carried out using the Symmetric Normalization algorithm. From the result of these steps, the volume is normalized to the atlas. The hippocampus segmentation from the atlas is then warped back to the input volume. Based on multiple atlases and a statistical learning method, the final segmentation is obtained.

AutoSeg[4] is able to do tissue-segmentation, parcellation, and segmentation of sub-cortical structures. After a bias field correction step, a rigid registration to a common coordinate system is carried out. Then, by using an expectation maximization segmentation algorithm, a tissue segmentation is carried out (into white matter, gray matter, and cortical spinal fluid). After a skull stripping step, the atlas is registered to the volume using a deformable registration. This allows

[1] Available at `http://www.nitrc.org/projects/hippseg_2011`
[2] version 5.1.0, available at `http://surfer.nmr.mgh.harvard.edu`
[3] version 1.0, available at `http://www.nitrc.org/projects/ahead`
[4] version 2.9, available at `http://www.nitrc.org/projects/autoseg`

to obtain the final segmentation based on the registration transform and the labels stored in the atlas.

Although **BrainParser**[5] is usually able to label various different subcortical structures, we use a version which is specifically tailored to hippocampus segmentation. After re-orienting the input volume to the coordinate system of the included, pre-trained atlas, skull stripping is performed. This is followed by computing an affine transform between the input volume and the reference brain volume. Then a deformable registration between the input and the reference volume is carried out. Next, according to the trained atlas, the input volume is labeled.

In case of BrainParser and AHEAD the MNI152 atlas has been used as provided with the software. For FreeSurfer and AutoSeg we used the MNI305 atlas and the UNC Adult Brain Atlas, respectively.

2.3 Usability and Time-Complexity of the Software Packages

All tested programs are composed of several sub-components which in many cases are not developed by the authors but reused from other tools, most notably tools from the ITK library. In order for the programs to be working correctly those sub-components have to work. In case of Brain Parser this means that three different versions of the ITK library (namely version 3.2, 3.16 and 3.20) have to be available because each tool is linked against a different ITK version.

Another problem is the general usability. The main executable of AutoSeg is a graphical user interface which seems rather user friendly at the first glance. However, in the background this tool just creates a script which executes all the steps necessary for the segmentation. The segmentation itself is again done by individual programs. However, the user interface gives no feedback on problems with the configuration or with the input data. Thus, detecting problems and their eventual causes is only possible by a time consuming investigation of logging information.

The best out-of-the box experience is provided by AHEAD and FreeSurfer.

For execution time-complexity, all packages were evaluated on a Linux system using an Intel® Core™2 CPU running at 2.66GHz. The fastest package was BrainParser with an average runtime per subject of about 2 hours followed by AutoSeg with an average runtime of about 6 hours. By far the slowest tools are AHEAD and FreeSurfer with a runtime of approximately 3 days and 1 day per subject, respectively. Although, in case of FreeSurfer it must be noted that the program performs a labeling of various different subcortical structures while BrainParser and AHEAD focus on the hippocampi only. AutoSeg also yields more structures than BrainParser and AHEAD but by far less as compared to FreeSurfer.

[5] available at `http://www.nitrc.org/projects/brainparser`

3 Segmentation Results

3.1 Metrics Used to Assess the Segmentation Quality

To allow assessing the quality of the automated hippocampus segmentation methods, metrics are needed.

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.

- **Similarity index (SI)**
  The similarity index (also known as the Dice coefficient) is a quite frequently used measure to assess the similarity between two sets of voxels.

$$SI(G, S) = \frac{2v(G \cap S)}{v(G) + v(S)} \tag{1}$$

- **Hausdorff distance (HD)**
  This metric is based on the actual structure of a voxel volume. It is defined as

$$HD(S, G) = \max_{x \in S}(\min_{y \in G} d(x, y)), \tag{2}$$

  where x and y are vectors in $\mathbb{R}^3$ and $d(\cdot, \cdot)$ denotes the Euclidean distance between two vectors.

In case of the similarity index a value of 1 denotes a perfect correspondence between G and S, whereas a value of 0 means that the intersection between the automated segmentation result and the ground truth is empty. The Hausdorff distance yields a value of 0 in case of a perfect correspondence between G and S. The higher the dissimilarity between G and S, the higher the respective distance value (there exists no upper bound for the values).

3.2 Experimental Segmentation Results

Results are based on both hippocampi simultaneously, viewing them as a single volume. Figures 1 and 2 show the similarities between the segmentations of the different programs and the ground truth created by one rater (and verified by two other raters, as provided with the dataset).

The results based on the the similarity index are quite inhomogeneous. In many cases we obtained the highest scores for AHEAD. But there are also cases where AutoSeg or FreeSurfer yield the highest scores. Only BrainParser seems to consistently deliver the lowest scores. In fact for 7 subjects BrainParser produces significantly lower scores (subjects 3, 14, 18, 21, 22, 23, and 24) and fails to produce a sensible segmentation for subject 8. AutoSeg fails at subjects 18 and 19. For the latter three segmentations, the computed volumes have no intersection with the corresponding ground truth volumes. Therefore, the similarity has to be zero according to the definition of SI. As observable below, the Hausdorff distance delivers (high) numerical values for these results since a volume intersection is not required for a valid result.

The Hausdorff distances in Fig. 2 again show that BrainParser consistently yields the highest distances (i.e. the lowest similarities). Also AutoSeg yields rather high scores (i.e. poor quality) for some subjects.

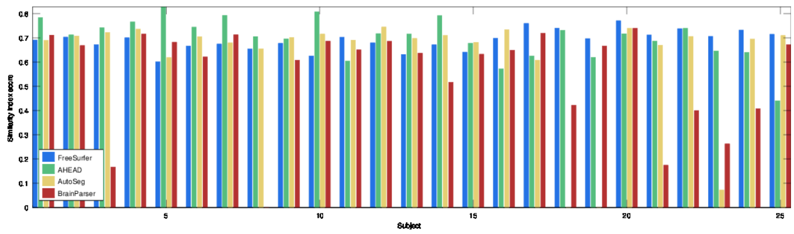

Figure 1. Comparison of automated segmentations for the Radiology Research database (based on the similarity index).

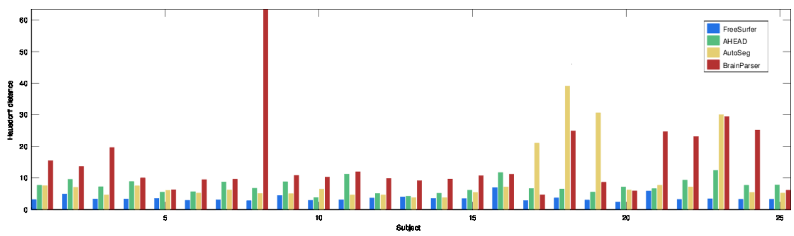

Figure 2. Comparison of automated segmentations for the Radiology Research database (based on the Hausdorff distance).

Table 1 summarizes the results (i.e. the means and standard deviations are computed over all subjects). While in case of the similarity index AHEAD seems to deliver the most accurate segmentations with respect to the ground truth, the most accurate program in case of the Hausdorff distance is FreeSurfer. But according to both metrics BrainParser (and AutoSeg) deliver the lowest overall accuracies (low quality segmentations also exhibited by significantly higher standard deviations, especially for BrainParser).

4 Conclusion

Two aspects of out-of-the-box software for hippocampus segmentation have been assessed in this work: Usability and segmentation accurracy. As discussed, with

Table 1. Summary of the results from the comparisons between the programs and the ground truth for the Radiology Research database (25 subjects).

	SI	**HD**
FreeSurfer	0.69±0.04	3.58±1.01
AHEAD	0.70±0.09	7.41±2.23
AutoSeg	0.62±0.23	9.68±9.56
BrainParser	0.55±0.21	15.33±12.21

respect to the first issue, AHEAD and Freesurfer are preferable over the other two software packges. Also with respect to segmentation accuracy, these two packages deliver the most accurate and reliable (with respect to complete segmentation failures and result variance) results. Since FreeSurfer tends to oversegmentations [8], a fusion of the output of those two packages might be a viable option to achieve accurate automated segmentation with controlled oversegmentation extent.

References

1. Fotuhi M, Do D, Jack C. Modifiable factors that alter the size of the hippocampus with ageing. Nat Rev Neurol. 2012;8:189–202.
2. Malmgren K, Thom M. Hippocampal sclerosis - origins and imaging. Epilepsia. 2012;53:19–33.
3. Teipel S, Grothe M, Lista S, Toschi N, Garaci F, Hampel H. Relevance of magnetic resonance imaging for early detection and diagnosis of Alzheimer disease. Med Clin North Am. 2013;97:399–424.
4. Fischl B, van der Kouwe A, Destrieux C, Halgren E, Ségonne F, Salat DH, et al. Automatically Parcellating the Human Cerebral Cortex. Cerebral Cortex. 2004;14(1):11–22.
5. 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. .
6. 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.
7. Jafari-Khouzani K, Elisevich K, Patel S, Soltanian-Zadeh H. Database of magnetic resonance images of nonepileptic subjects and temporal lobe epilepsy patients for validation of hippocampal segmentation techniques. Neuroinformatics. 2011;9(4):335–346.
8. 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.