Auto-kNN: Brain Tissue Segmentation using Automatically Trained k-Nearest-Neighbor Classification

Henri A. Vrooman[1], Fedde van der Lijn[1] and Wiro J. Niessen[1,2]

[1] Biomedical Imaging Group Rotterdam, Departments of Medical Informatics and Radiology, Erasmus MC - University Medical Center Rotterdam, The Netherlands
[2] Imaging Science and Technology, Faculty of Applied Sciences, Delft University of Technology, The Netherlands

Abstract. In this paper we applied one of our regularly used processing pipelines for fully automated brain tissue segmentation. Brain tissue was segmented in cerebrospinal fluid (CSF), gray matter (GM) and white matter (WM). Our algorithms for skull stripping, tissue segmentation and white matter lesion (WML) detection were slightly adapted and applied to twelve data sets within the MRBrainS13 brain tissue segmentation challenge. Skull stripping is performed using non-rigid registration of 5 atlas masks. Our tissue segmentation is based on an automatically trained kNN-classifier. Training samples were obtained by non-rigid registration of 5 manually labeled scans followed by a pruning step in feature space to remove any residual erroneously sampled tissue voxels. The kNN-classification incorporates voxel intensities from a T1-weighted scan and a FLAIR scan. The white matter lesion detection is based on an automatically determined threshold on the FLAIR scan. The application of the algorithms on the data from the MRBrainS13 Challenge showed that our pipeline produces acceptable segmentations. Average resulting Dice scores were 77.86 (CSF), 81.22 (GM), 87.27 (WM), 93.78 (total parenchyma), and 96.26 (all intracranial structures). Total processing time was about 2 hours per subject.

Keywords: Brain tissue segmentation, kNN-classifier, Automated classifier training, White matter lesion detection, Skull stripping

1 Introduction

The segmentation of magnetic resonance (MR) images in white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF) is one of the classic neuroimage analysis challenges. Brain tissue volume measurements are used in studies on ageing and neurodegenerative diseases [1,2]. These segmentations are also commonly employed as regions of interest for other neuroimage analyses, including cortical thickness measurements [3], voxel-based morphometry [4], and connectivity analyses [5].

Given the relevance of brain tissue segmentation, many automated different segmentation methods have been proposed over the years. Almost all of these methods rely on a supervised or unsupervised voxel classifier. Supervised methods use manu-

adfa, p. 1, 2011.

ally segmented training data to learn the typical distribution of intensity or appearance features for the tissue classes [6]. This has the advantage that the classifier explicitly follows the manual segmentation protocol and it allows the use of large amounts of features. Supervised methods, however, also require that the training data resembles the unlabeled target scan. Since the manual segmentation of a few MR images is very time- and labor-intensive suitable training data is not allows available.

Unsupervised methods, particularly those based on expectation maximization (EM), do not require training data and are therefore more widely used than supervised methods. EM-based methods start with an initial segmentation, which is often based on a probabilistic brain tissue atlas that is registered to the unlabeled target scan. From this initialization, class-specific Gaussian intensity distributions are estimated. This intensity model can then be used to update the segmentation and this process is repeated until the segmentation converges.

The main reason for its popularity is that several EM-based methods are publicly available [7,8], as standalone application or as part of a neuroimaging analysis package like SPM [9,10] and FSL [11]. But the EM framework is also attractive from a methodological perspective. In particular, the segmentation procedure can be easily enhanced with features like bias field estimation, Markov Random Field (MRF) regularization, and an iteratively updated atlas registration.

In this paper we validate an alternative fully automated and unsupervised segmentation method, originally introduced in Cocosco et al. [12] and later expanded in Vrooman et al. [13] and De Boer et al. [14]. This automatically trained k-Nearest-Neighbors segmentation method (auto-kNN) uses a similar strategy of initialization, estimation, and segmentation but offers two advantages over the previously mentioned EM-based methods: it uses a non-parametric kNN-classifier that can model more complex decision boundaries than a Gaussian intensity model; and it uses multi-atlas registration to estimate the intensity model which is known to be more accurate and robust than single atlas registration [15].

2 Methods

The core of the method used in this Grand Segmentation Challenge was previously described in De Boer et. al. [14]. This section will provide a summary of this technique. We did make some minor changes in the pre- and post-processing, which will also be outlined below.

2.1 Data

All experiments were performed on multi-sequence MR imaging data made available for this challenge. The scans were acquired at the UMC Utrecht (the Netherlands) in patients with diabetes and matched controls (with increased cardiovascular risk). In this work we used the following sequences: a T1-weighted scan (T1w), a T1-weighted inversion recovery scan (IR), and a fluid attenuation inversion recovery scan (FLAIR). All images had a voxel size of 0.958mm x 0.958mm x 3.0mm and were

corrected for MR bias field artifacts. The IR and FLAIR images were also co-registered to the T1w.

Five manually segmented datasets were available for training and parameter tuning. Labels were provided for the background, cortical GM, basal ganglia, WM, white matter lesions (WML), cortical CSF, ventricular CSF, cerebellum, and the brain stem. Twelve datasets were supplied to test the proposed method. For these datasets manual labels were held back for unbiased testing. Segmentations were evaluated on three tissue classes GM (cortical GM and basal ganglia), WM (WM and WMLs), and CSF (cortical CSF, cistern CSF and ventricles). Cerebellum and brainstem voxels were ignored during validation. All manual segmentations were performed by one of two trained observers on the T1w using a contouring tool.

2.2 Preprocessing

Preprocessing consisted of two steps: masking and intensity normalization. For the training images masks were obtained by binarizing the manual segmentations. For the test images, masks were computed using a multi-atlas segmentation method. As atlases we used the T1w training images, both in the original and in a left-right-flipped version, and their associated brain masks. Each atlas image was registered to the unlabeled test images using Niftyreg [16]; the registration was applied by computing an affine transformation, followed by a non-rigid deformation (using a 5mm B-spline grid and normalized mutual information). A final mask was then computed using STEPS [17]. This method deforms both atlas images and labels, selects per voxel location the five most similar atlases (based on local normalized cross correlation), and fuses their labels using STAPLE [18]. The masking procedure was different from De Boer et al. [14] in which a single atlas was used to obtain the intracranial space. Intensity normalization was performed for all images by a linear mapping obtained by setting the lower and upper 4th percentile intensities to zero and one, and interpolating the values in between.

2.3 Tissue segmentation

The tissue segmentation is initialized by constructing GM, WM, and CSF tissue probability maps in the unlabeled target image coordinate frame using multi-atlas registration. These maps are then used as a mask to sample multi-modal intensities from the target image. To correct for residual misregistration a pruning operation is performed using a clustering method. Finally, the resulting target-specific samples are used to train a kNN classifier that can be used to segment the target image.

For the challenge data, the tissue probability maps were constructed using the five manually labeled T1w training images. GM, WM, and CSF segmentations were created by fusing the eight available labels as described above. The atlases were first affinely registered to the target images, followed by a non-rigid B-spline registration. The deformations were computed with Elastix [19] using mutual information as similarity measure and a 5 mm B-spline grid. The tissue probability maps were then obtained by deforming the atlas labels and averaging them.

Multi-dimensional intensity brain tissue samples were then extracted from all images of the each target subject by thresholding the GM, WM, and CSF probability maps at 0.7 and randomly choosing 7500 voxels per class. These settings were based on previously published parameter tuning experiments on different data [13, 14]. This procedure allowed us to benefit from the well-documented ability of multi-atlas registration to compensate for registration errors. To remove any residual erroneously sampled tissue voxels a pruning step was performed. This was done by mapping the samples in the multi-dimensional feature space defined by the intensities of all scans and computing a minimum spanning tree. The tree was then iteratively pruned by removing connections for which its length exceeds a threshold of a constant times the average length of the other connections of a sample. This process is repeated until three large clusters remain that predominantly contain a single tissue class. All minority tissue samples in these three clusters were then removed, as well as any unconnected smaller clusters.

After this pruning step a k-Nearest-Neighbor classifier with a k-value of 45 was trained in the same feature space as the minimum spanning tree. The value for k was again based on previous experiments [13]. Finally, a segmentation was obtained by applying this classifier to the target subject images.

2.4 White matter lesion segmentation and post-processing

The white matter lesion detection step was based on an automatically-selected threshold on the FLAIR scan [14]. First the tissue segmentation from the previous step was used to localize the GM voxels in the FLAIR scan. Assuming that the voxels with the highest intensity in the histogram of these voxels are WML candidates, a threshold was set on 2.3 standard deviations higher than the location of the top of the smoothed histogram. This parameter was set based on previous experiments. The WML segmentation was further refined using two minor morphological operations [14].

Based on the visual inspection of the test results, two ad-hoc morphological operations were included to refine the results. Firstly, small local minima were relabeled as the surrounding class (MevisLab®: itkGrayscaleFillholeImageFilter), especially to fill small areas (< 3x3x3 voxels) in CSF that were labeled as background voxels. Since the brain masks had the tendency to overestimate the intracranial space, they included parts of the dura or bone marrow that were labeled as GM. Therefore, we applied a second post-processing step, in which GM voxels that directly bordered the background were relabeled as background. These operations were both not included in the work of de Boer et al. [14].

3 Experiments and results

3.1 Parameter tuning

To tune our processing pipeline for the challenge date, we created brain masks for the five training subjects by binarizing the corresponding label images (CSF+GM+WM) as said in Section 2.2. The probability maps for background, CSF,

GM and WM were created using a non-rigid registration (as described in Section 2.3) of the label images using a leave-one-out-procedure. For each of the five training subjects the label images of the other four were used as atlases.

For the application of our automatically trained kNN-classifier, we tested different combinations of input sequences. In Table 1, the resulting Dice factors for three combinations are shown: 1) T1w + IR; 2) T1w + FLAIR; and 3) T1w + FLAIR +IR. Since the combination of the T1w and FLAIR scans yielded the best scores, we used that combination to segment the test subjects.

Input Sequences	**Subj. 1**	**Subj. 2**	**Subj. 3**	**Subj. 4**	**Subj. 5**
T1w + IR					
CSF Dice (%)	86.71	80.54	82.58	88.14	80.66
GM Dice (%)	81.74	77.89	80.94	83.14	84.43
WM Dice (%)	86.18	80.84	87.80	86.75	88.92
T1w + FLAIR					
CSF Dice (%)	87.40	87.86	86.88	88.89	87.30
GM Dice (%)	80.93	82.59	81.72	83.78	86.18
WM Dice (%)	85.95	83.45	88.21	87.16	90.28
T1w + IR + FLAIR					
CSF Dice (%)	84.91	86.23	84.27	88.27	85.91
GM Dice (%)	80.69	81.33	81.70	83.73	86.14
WM Dice (%)	86.73	81.84	87.61	86.69	89.75

Table 1. The resulting Dice factors (%) for the segmentation of CSF, GM, and WM in the training subjects, using different combinations of input scans (T1w + IR, T1w + FLAIR, and T1w + IR + FLAIR) for the kNN-classifier. With the second combination (T1w + FLAIR) we obtained the highest Dice scores during tuning.

As a second tuning step, we varied a number of parameters for the kNN-classifier. However, changing the number of samples, the k-value or the threshold for the probability maps didn't change the outcome scores in a significant way. We decided to use the default values as described in Section 2.3. Also for the WML detection, further tuning of the parameters did not improve the training results. In our opinion, this shows that our WML detection and our automatically trained kNN-classifier are reasonably robust when applied to different cohorts.

3.2 Challenge data segmentation

We applied the method to the twelve test images. Results were visually inspected and compared to the manual segmentations by the Challenge organizers. The accuracy was evaluated using the following measures:

- Dice coefficient (DC) [20]
- Modified Hausdorff distance (MHD) [21]
- Absolute volumetric difference (AVD) [22]

In Table 2, the mean and standard deviation (Std) of the final scores for the twelve test subjects, as computed by the Challenge organizers, are shown. Five tissues were evaluated: GM, WM, CSF, brain (WM + GM), and all intracranial structures (WM + GM + CSF). From the three brain tissue classes, WM scored the best. Overall, CSF had the lowest evaluation scores.

	Dice (%)		**HD (mm)**		**AVD (%)**	
Structure	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**
Gray matter	81.22	1.80	3.86	0.88	6.58	4.53
White matter	87.27	0.96	3.02	0.41	7.57	4.21
Cerebrospinal fluid	77.86	4.98	3.21	0.64	17.88	15.73
Brain	93.78	0.75	4.94	1.31	3.93	2.41
All intracranial structures	96.26	1.26	3.99	1.04	3.79	2.95

Table 2. The resulting scores for the twelve challenge training subjects, as calculated by the MRBrainS13 Challenge Board.

In Figure 1, the tissue segmentation for test subject 3 is shown including the WML detection. Subject 3 has a large number of white matter lesions and Figure 1 shows that our WML detection is able to detect the white matter lesions in an acceptable way. For this challenge, only CSF, GM, and WM were required. Therefore, we relabeled the white matter lesions to WM to obtain the final segmentation results.

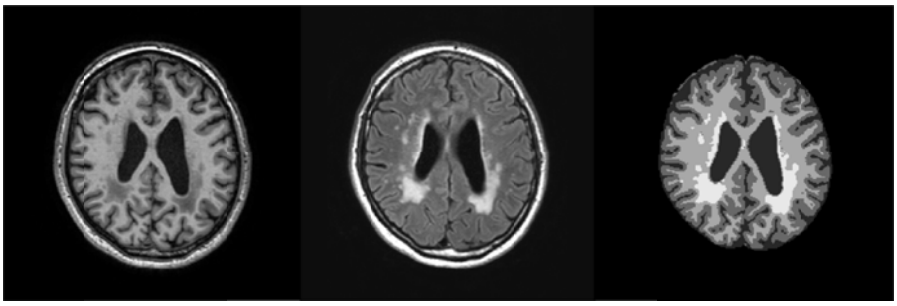

Fig. 1. Segmentation result for Subject 3 (slice 29). Three data sets are shown: the T1w scan (left), the FLAIR scan (middle) and the final segmentation including the detection of WML (right). Subject 3 had a large number of relatively large WML. The result shows that our pipeline is able to detect the WML with high accuracy.

In Figure 2, the resulting tissue segmentation for three test subjects are shown. We noticed by visual inspection that probably two aspects had a slightly negative influence on our results. In the first place, since the brain masks had the tendency to overestimate the intracranial space, it seems that in some cases there is too much CSF labeled in the upper-front part of the intracranial space. Furthermore, parts of the dura or bone marrow were labeled as GM.

As described in Section 2.4, we applied two extra post-processing steps to diminish these artifacts. These steps improved the segmentations a lot, but it was not possible to cover this issue completely with the currently applied processing pipeline. Secondly, in some cases the thalamus and the globus pallidus are not segmented completely into GM. This is a well-known limitation of our current processing pipeline and can in our opinion be improved in the future by using the registration of appropriate atlases to the subject of interest.

The total runtime of our pipeline was about 2 hours per test subject. The creation of the mask needed 10 non-rigid registrations (8 minutes each). Registration of the 5 manual labeled images followed by the computation of the probability maps took 30 minutes. The application of the kNN-classifier, the WML detection, and the post-processing steps were done in about 5 minutes. The registrations for the image mask and probability maps were computed on a cluster with AMD Opteron 2216 2.4GHz nodes without multi-threading, the kNN-classifier, the WML detection and the post-processing were implemented in MevisLab® and computed on a machine with an Intel® Xeon® E5620 2.40 GHz CPU, 12 GB of installed memory and 64-bit Windows 7 Operating System.

4 Discussion

In this paper, we showed the results of our brain tissue segmentation pipeline, applied to twelve subjects within the MRBrainS13 tissue segmentation challenge. We applied one of our regularly used processing pipelines. We have presented an algorithm for automated brain extraction and brain tissue segmentation. The skull stripping algo-

rithm is based on multi-atlas segmentation utilizing the STEPS [14] algorithm; for tissue classification we used an automatically trained kNN-classifier with voxel intensities obtained from the T1w scan and FLAIR scan as input features. WML detection was based on an automatically set threshold on the intensities in the FLAIR scan.

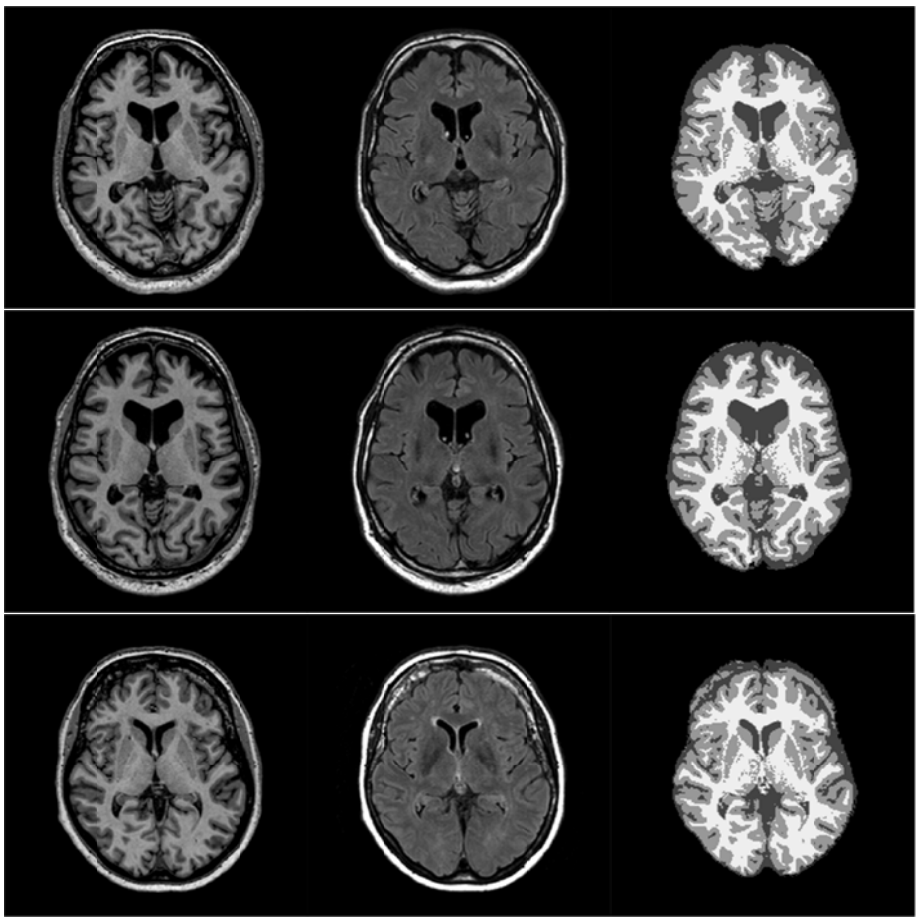

Fig. 2. Segmentation results for three subjects (8, 9 and 10). For all subjects slice 21 is shown above. Three data sets are shown: the T1w scan (left), the FLAIR scan (middle) and the final segmentation (right). Subject 8 (top) had the highest average scores. Subject 9 (middle) is an average result. Subject 10 (bottom) had the lowest evaluation scores. The misclassification in the frontal CSF and the under-segmentation of the thalamus are visible especially in Subject 10. The under-segmentation of the globus pallidus is visible in all subjects.

The main advantage of our approach, is that training data is not needed for tissue classification. The training samples are extracted from the subject data itself, after

non-rigid registration of brain tissue atlases. Therefore, our segmentation pipeline works robust and accurate on data obtained from different scanners and with different scan protocols. Our in-house tissue atlases can be used for several, different populations. We have noticed, however, that using atlases matching the population better gives slightly better results.

In our opinion, we obtained reasonable segmentation results on the challenge data, with a slightly adapted pipeline. The segmentation results are sometimes a little bit more noisy than segmentation results published elsewhere, since we don't use for example a Markov Random Field or morphologic post-processing steps (smoothing of kNN posteriors) to clean up or smooth the segmentation result. We are planning to incorporate this step in the future, although we expect volume measures not to be influenced significantly by such a processing step.

Another issue is that our algorithm sometimes does not detect the total amount of GM in the globus pallidus and thalamus as can clearly be seen in Figure 2. This is a well-known problem with brain tissue segmentation. The globus pallidus derives its name from its pallid appearance in fresh unstained specimen. It is traversed by numerous bundles of heavily myelinated fibers, which give a relatively high intensity on a T1w scan and distinguish it from the dark appearance of for example the corpus striatum, the putamen and the caudate nucleus, which are generally better segmented. We have seen in previous cases that other type of scans, i.e. a T2w scan can improve the segmentation of the basal ganglia. Another possibility is the registration of atlases containing the manual segmentation of specific brain structures.

Looking at the final evaluation scores delivered by the challenge board, it is obvious that most errors were made on the CSF. In our opinion, this is mainly due to the fact that the created brain mask was sometimes crossing the intracranial border due to slight misregistration of the mask atlases. In previous applications of our segmentation pipeline only the cerebrum was involved. Since for this challenge also the CSF around cerebellum and brain stem had to be segmented, we needed to create and register new brain masks. The skull stripping is in our opinion still one of the most difficult steps in segmentation pipelines in general, leading to subsequent errors in the classification process. The applied post-processing steps could improve our results but not completely eliminate the errors. Our evaluation scores for intracranial volume are reasonable, for example compared to a publication of Iglesias [23]. Since the CSF, however, is a relatively small set of voxels within the mask, the influence of mask errors on CSF segmentation is substantial. We are planning to improve the skull stripping step in the near future to deal with this overestimation of CSF.

In this challenge, a set of five training sets (including labels) were made available. We like to mention that if no manual segmentations are at hand, our processing pipeline can also be used using other atlases obtained from other institutions or build in-house. In the last few years we are applying our brain segmentation pipeline on different cohorts (scanned on several types of scanners and with varying scan protocols). In most cases we achieve acceptable results using our in-house set of brain atlases.

The total runtime of our algorithm was about 2 hours per subjects. Most of the time was needed for the 15 non-rigid registrations involved. We are currently working on a parallelization of the Elastix code to reduce processing times. Another possibility

to increase computation speed is the integration of the registration needed for skull stripping and for the creation of tissue probability maps. Speeding up the processing pipeline is especially relevant for the translation of our processing pipeline to a radiologic workstation in the clinic. For large population studies, increasing the speed of the pipeline is less urgent, since in that case we process large cohorts of patients or healthy subjects in parallel on a multi-core computing cluster.

We believe that if the limitations mentioned above are further investigated and solved, our fully automated brain tissue segmentation pipeline can be applied to a diverse set of cohorts with an accuracy and reproducibility that are comparable human raters.

References

1. Fotenos, A.F., Snyder, A.Z., Girton, L.E., Morris, J.C., Buckner, R.L.: Normative estimates of cross-sectional and longitudinal brain volume decline in aging and AD. Neurology **64** (2005) 1032–9
2. Ikram, M.A., Vrooman, H. a, Vernooij, M.W., van der Lijn, F., Hofman, A., van der Lugt, A., Niessen, W.J., Breteler, M.M.B.: Brain tissue volumes in the general elderly population. The Rotterdam Scan Study. Neurobiology of aging **29** (2008) 882–90
3. Good, C.D., Johnsrude, I.S., Ashburner, J., Henson, R.N., Friston, K.J., Frackowiak, R.S.: A voxel-based morphometric study of ageing in 465 normal adult human brains. NeuroImage **14** (2001) 21–36
4. Kim, J.S., Singh, V., Lee, J.K., Lerch, J., Ad-Dab'bagh, Y., MacDonald, D., Lee, J.M., Kim, S.I., Evans, A.C.: Automated 3-D extraction and evaluation of the inner and outer cortical surfaces using a Laplacian map and partial volume effect classification. NeuroImage **27** (2005) 210–21
5. Robinson, E.C., Hammers, A., Ericsson, A., Edwards, A.D., Rueckert, D.: Identifying population differences in whole-brain structural networks: a machine learning approach. NeuroImage **50** (2010) 910–9
6. Anbeek, P., Vincken, K.L., van Bochove, G.S., van Osch, M.J.P., van der Grond, J.: Probabilistic segmentation of brain tissue in MR imaging. NeuroImage **27** (2005) 795–804
7. Cardoso, M.J., Leung, K., Modat, M., Keihaninejad, S., Cash, D., Barnes, J., Fox, N.C., Ourselin, S.: STEPS: Similarity and Truth Estimation for Propagated Segmentations and its application to hippocampal segmentation and brain parcelation. Medical image analysis **17** (2013) 671–684
8. Cardoso, M.J., Clarkson, M.J., Ridgway, G.R., Modat, M., Fox, N.C., Ourselin, S.: LoAd: a locally adaptive cortical segmentation algorithm. NeuroImage **56** (2011) 1386–97
9. Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P.: Automated model-based bias field correction of MR images of the brain. IEEE Trans. Med. Imag. **18** (1999) 885–96
10. Ashburner, J., Friston, K.J.: Unified segmentation. NeuroImage **26** (2005) 839–51
11. Zhang, Y., Brady, M., Smith, S.: Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med Imag. **20** (2001) 45–57
12. Cocosco, C.A., Zijdenbos, A.P., Evans, A.C.: A fully automatic and robust brain MRI tissue classification method. Medical image analysis **7** (2003) 513–27

13. Vrooman, H.A., Cocosco, C.A., van der Lijn, F., Stokking, R., Ikram, M.A., Vernooij, M.W., Breteler, M.M.B., Niessen, W.J.: Multi-spectral brain tissue segmentation using automatically trained k-Nearest-Neighbor classification. NeuroImage **37** (2007) 71–81
14. de Boer, R., Vrooman, H.A., van der Lijn, F., Vernooij, M.W., Ikram, M.A., van der Lugt, A., Breteler, M.M.B., Niessen, W.J.: White matter lesion extension to automatic brain tissue segmentation on MRI. NeuroImage **45** (2009) 1151–61
15. Heckemann, R.A., Hajnal, J. V, Aljabar, P., Rueckert, D., Hammers, A.: Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage **33** (2006) 115–26
16. Modat, M., Ridgway, G.R., Taylor, Z.A., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., Ourselin, S.: Fast free-form deformation using graphics processing units. Computer methods and programs in biomedicine **98** (2010) 278–84
17. Cardoso, J. M., Leung K., Modat M., Keihaninejad S., Cash D., Barnes J., Fox N.C., Ourselin S.: STEPS: Similarity and Truth Estimation for Propagated Segmentations and its application to hippocampal segmentation and brain parcelation. Med. Image Anal. **17**(6) (2013) 671-84
18. Warfield S.K., Zou K.H., and Wells W.M.: Simultaneous truth and performance level estimation (STAPLE): An algorithm for validation of image segmentation, IEEE Trans. Med. Imag. **23**(7) (2004) 903-921
19. Klein S., Staring M., Murphy K., Viergever M.A., and Pluim, J.P.W.: Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imag. **29** (2010) 196-205
20. Dice, L.: Measures of the amount of ecologic association between species. Ecology **26**(3) (1945) 297–302
21. Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J.: Comparing images using the Hausdorff distance. IEEE Transactions on Pattern Analysis and Machine Intelligence **15**(9) (1993) 850–863
22. Babalola, K.O., Patenaude, B., Aljabar, P., Schnabel, J., Kennedy, D., Crum, W., Smith, S., Cootes, T., Jenkinson, M., Rueckert, D.: An evaluation of four automatic methods of segmenting the subcortical structures in the brain. Neuroimage **47**(4) (2009) 1435–1447
23. Iglesias, J.E., Liu, C.-Y., Thompson, P.M., Tu, Z.: Robust brain extraction across datasets and comparison with publicly available methods. IEEE trans. Med. Imag. **30** (2011) 1617–34