Segmentation of Subcortical structures in T1 weighted MRI as a component of a Brain Atrophy Computation Pipeline

Master Thesis

Cecilie Benedicte Anker

Biomediq A/S
&
Technical University of Denmark, DTU

Supervised by:
Prof. Mads Nielsen, KU, Prof. Rasmus Larsen, DTU,
Prof. Knut Conradsen, DTU, & Postdoc Mark Lyksborg, DTU.

2014

Technical University of Denmark
DTU Compute
Building 303B, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253031, Fax +45 45881399
reception@compute.dtu.dk
www.compute.dtu.dk

Abstract

Among the top performing automated hippocampal segmentation methods from structural Magnetic Resonance Imaging (MRI), are multi-atlas segmentation methods, which rely on manual annotations.

In this thesis two fundamentally different multi-atlas segmentation methods are implemented, *N-L Patch* and *BrainFuseLab*. In N-L Patch, each voxel is segmented using information from atlases which have been coarsely aligned using affine registrations. BrainFuseLab aligns atlases using non-rigid registrations, and is thus comparatively slower. To make a fair comparison, both methods will use the same atlases from a new Harmonized Hippocampal Protocol (HHP).

Method parameters are optimized in a leave-one-out cross-validation using two different atlas sets. Based on volume overlap with the manual annotations, N-L Patch is chosen to segment a standardized ADNI dataset containing 1.5T MRIs from 504 diagnosed subjects (169 cognitively normal (CN), 234 mild cognitive impairment (MCI), 101 alzheimer's disease (AD)) at baseline, month 12 and month 24. Hippocampal atrophy calculated as percentage volume change from baseline to follow-up is estimated. Based on a statistical analysis, the diagnostic group separation capabilities of N-L Patch are compared to two state-of-the-art methods, cross-sectional FreeSurfer and longitudinal FreeSurfer.

Including the HHP annotations in N-L Patch yielded significantly better group separation than cross-sectional FreeSurfer in separating AD from CN and AD from MCI. This illustrates the longitudinal robustness of segmentations when annotations from the new hippocampal standard are included in automated segmentation methods. Also longitudinal FreeSurfer exploiting baseline and follow-up simultaneously showed no diagnostic improvement over N-L Patch.

Resumé

Multi-atlas segmenteringsmetoder med manuelle annoteringer er blandt de bedste automatiske hippocampus segmenteringsmetoder til strukturel Magnetisk Resonans (MR) billeder.

I denne afhandling er to fundamentalt forskellige multi-atlas segmenteringsmetoder implementeret, *N-L Patch* og *BrainFuseLab*. I N-L Patch er hver voxel segmenteret ved at bruge information fra atlaser, der er groft rettet ind ved affine registreringer. BrainFuseLab retter atlaserne ind ved brug af ikke-rigide registreringer og er derfor relativt langsommere beregningsmæssigt. Begge metoder benytter de samme atlaser fra en ny Harmoniseret Hippocampus Protokol (HHP).

Metodeparametre er optimerede i en *leave-one-out* krydsvalidering. Baseret på volumenoverlap med de manuelle annoteringer er N-L Patch valgt til at segmentere et standardiseret ADNI datasæt der indeholder 1.5T MR billeder fra 504 diagnostiserede forsøgspersoner (169 kognitiv normal (CN), 234 mild kognitiv forringelse (MCI), 101 alzheimers sygdom (AD)) ved udgangspunktet, måned 12 og måned 24. Hippocampusatrofi, beregnet som den procentvise forskel fra udgangspunktet til opfølgning, er estimeret. Ud fra en statistisk analyse, er N-L Patchs diagnostiske separationsevne sammenlignet med to *state-of-the-art* metoder, *cross-sectional FreeSurfer* og *longitidinal FreeSurfer*.

Ved at inkludere HHP annoteringer i N-L Patch fås signifikant bedre adskillelse af AD fra CN og AD fra MCI end for *cross-sectional FreeSurfer*. Dette illustrerer segmenteringsrobusthed over tid når annoteringer fra den nye hippocampus standard inkluderes i automatiske segmenteringsmetoder. Også *longitudinal FreeSurfer*, der bruger information fra udgangspunkt og opfølgning samtidig, viste ingen forbedret diagnostisk separationsevne i forhold til N-L Patch.

Preface

This thesis was prepared at the Department of Applied Mathematics and Computer Science (DTU Compute) at the Technical University of Denmark (DTU) in partial fulfillment of the requirements for acquiring the Master of Science Degree in Engineering, M.Sc.Eng.
The undersigned is a Master student in Medicine & Technology at the Technical University of Denmark and the Faculty of Health Sciences, University of Copenhagen (KU).
The work was carried out at the company Biomediq A/S, Fruebjergvej 3, 2100 Copenhagen, Denmark. The thesis was supervised by Professor Rasmus Larsen (DTU), Professor Knut Conradsen (DTU), Postdoc Mark Lyksborg (DTU) and Professor Mads Nielsen (Biomediq A/S & KU). The work was carried out from September 2013 to February 2014, corresponding to 30 ECTS credits.

Copenhagen, February 2014

Cecilie Benedicte Anker

Acknowledgements

I would like to thank my DTU supervisors Professor Rasmus Larsen, Professor Knut Conradsen and Postdoc Mark Lyksborg for their inputs during the weekly Friday meetings. A special thanks to Mark Lyksborg for helping me combining registrations in SPM.

I thank Professor Mads Nielsen from Biomediq A/S for advise and suggestions during supervision meetings and weekly meetings in the Alzheimer's group at Biomediq A/S. I have very much appreciated working on equal terms with the other employees.

A special thanks to PhD student Akshay Pai and Postdoc Lauge Sørensen from Biomediq A/S for helping with practicalities of any kind including finding relevant literature and to introduce me to the cluster file system.

For helping me to choose appropriate methods to implement, I would like to thank Associate Professor Koen Van Leemput and PhD student Oula Puonti, DTU. Furthermore, I thank Oula Puonti for advise regarding installation of BrainFuseLab.

Contents

Chapter 1

Introduction

In the United States, 45% of all people above 85 years suffer from the most common form of dementia, Alzheimer's disease (AD). The prevalence increases with the average lifetime year by year.
Payments for AD patients care for 2012 were estimated to be $200 billion in the United States, but the amount is expected to increase to $1,1 trillion in 2050 (in 2012 dollars) if medication is not improved [20].

AD is pathologically characterized by the presence of intracellular neurofibrillary tangles made of tau protein, extracellular amyloid plaques and decreasing brain volume (atrophy) due to death of brain cells (neurons). The steadily decreasing number of neurons affect a persons behavior, memory and ability to think clearly. At some point, the brain changes impair the ability to carry out basic functions such as swallowing and ultimately AD is fatal. At the moment no cure for AD is on the market [20].
Figure 1.1 illustrates that deaths caused by AD have continued to rise, while other major causes of death have decreased in the past years. This clarifies the need for developing new medication, which can cure AD or significantly decrease the disease progression rate.

One of the subcortical brain structures showing early pathological atrophy in AD is hippocampus, which is associated with consolidation of information from short-term memory to long-term memory and spatial navigation.

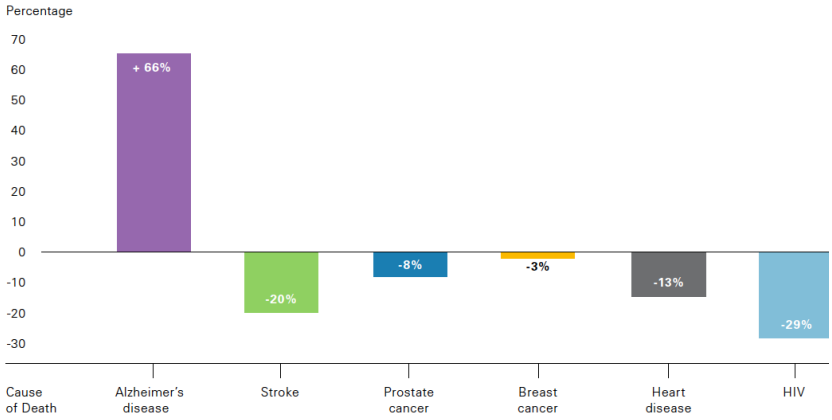


Figure 1.1: Percentage changes in selected Causes of Death (All ages) between 2000 and 2008 [20].

Hippocampal volumetry derived from structural Magnetic Resonance Imaging (MRI) has been endorsed by the new AD diagnostic guidelines as a radiological marker of disease progression [27] and proposed as a part of a new criteria to allow diagnosis of AD to be made earlier than it would be possible on pure clinical grounds [3]. Therefore, it is needed to segment the structure from T1-weighted MRI to analyze shape, volume and texture changes. A delimitation of hippocampus (blue) from MRI in a coronal, sagittal and transversal view can be seen in Figure 1.2. As the figure illustrates, a person has two hippocampi, one in each brain hemisphere.

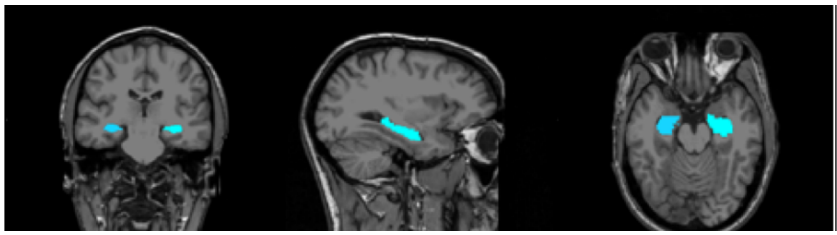
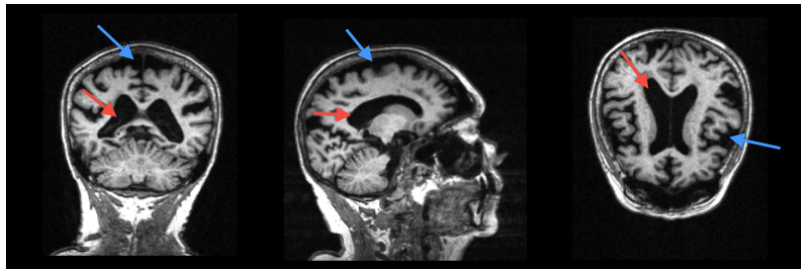


Figure 1.2: Segmentation of hippocampus (blue) from T1-weighted MRI, coronal, sagittal and transversal view.

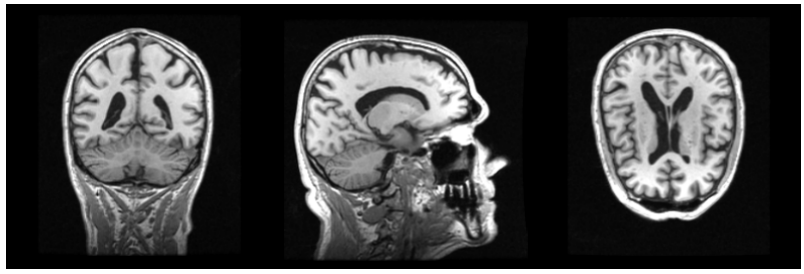
Manual segmentations of subcortical structures are very time consuming and are subject to errors [4]. To be practicable for studies with many subjects and in clinical applications, automated segmentation is needed [15]. This thesis will concern automated hippocampal segmentation from T1-weighted MRI.

1.1 Clinical background

In recent years, AD research has emphasized that decline in pathological processes and clinical functions occur gradually with dementia representing the end stage of many years of accumulation of these pathological changes. The pathological changes begin to occur decades before the earliest clinical symptoms [24]. The hypothesis is, that the pathological changes begin with abnormal processing of amyloid precursor protein (APP). APP leads to excess production or reduced clearance of β-amyloid (Aβ) in the cortex. Some of the Aβ-residues, especially Aβ_{42}, are highly hydrophobic and forms oligomers and fibrils, which accumulate as extracellular plaques. Furthermore, the Aβ oligomers lead to a cascade characterized by abnormal tau aggregation called neurofibrillary tangles (NFTs) inside the neurons, synaptic dysfunction, cell death, localized atrophy and eventually whole brain atrophy. Whole brain atrophy and enlarged ventricles are signs of AD progression and can be seen from MRI. In Figure 1.3 an AD and a normal aging subject's MRI can be seen. Both subjects are 84 years.



Alzheimer's Subject



Normal Aging

Figure 1.3: Coronal, sagittal and transversal view of a AD (row 1) and a normal aging (row 2) brain from T1-weighted MRI. Both subjects are 84 years. Blue arrows indicate whole brain atrophy and red arrows indicate enlarged ventricles.

In less than 1% of all who develop AD, the disease is caused by genetic mutations. In these cases disease symptoms tend to develop early, sometimes as early as age 30. In the more common form of AD called late-onset AD, symptoms normally occur at age 65 or older [20].

The clinical disease stages of AD can be divided into 3 stages. The first is a pre-symptomatic phase in which people are *Cognitively Normal*, CN. However, some have pathological changes in the brain. The second stage *Mild Cognitive Impairment*, MCI, is characterized by the onset of the earliest cognitive symptoms that do not meet the criteria of dementia. The third and final phase is AD dementia, defined as impairments in multiple domains that are severe enough to cause loss of function [24].

AD biomarkers, both chemical and imaging, do not peak simultaneously but rather in an ordered manner. Figure 1.4 illustrates the proposed dynamic view of AD in the forms of biomarkers, memory and clinical functions as a function of disease stage.

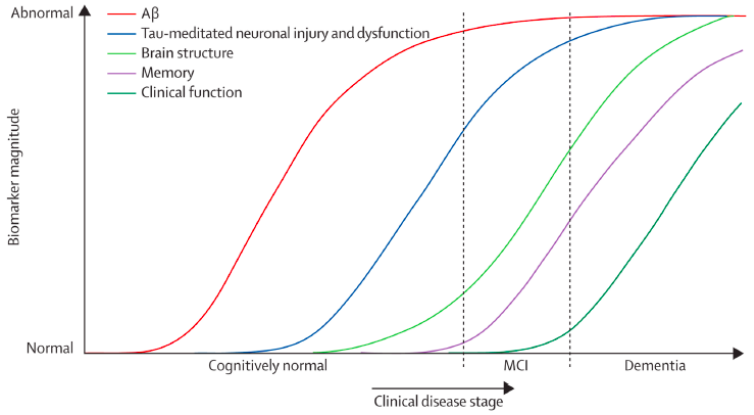


Figure 1.4: Dynamic view of AD biomarkers, memory and clinical function as a function of clinical disease stage [24].

Volumetric measures of brain atrophy show a strong correlation between the severity of atrophy and the severity of cognitive impairment in patients along the continuum from CN to AD. Hippocampus is in this context an interesting structure, because it is affected early and severely [26].

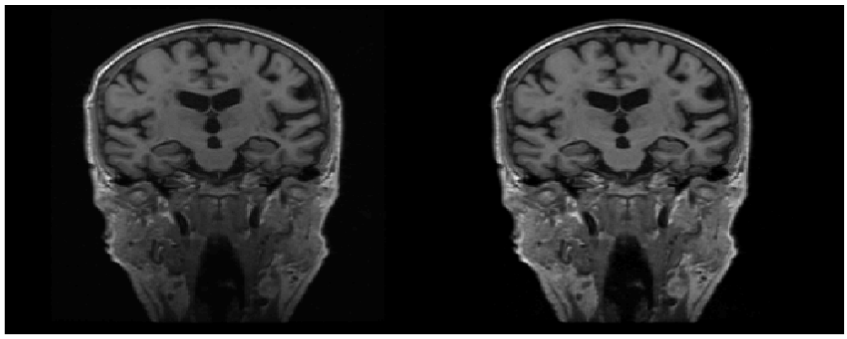
1.2 Current method

In recent years, many have developed automatic segmentation methods of structural T1-weighted MRI. A selection of these methods are explained in Chapter 2. One current and well-recognized method is *FreeSurfer*, [19], [4]. FreeSurfer is used at the company *Biomediq A/S* to segment subcortical brain structures including hippocampus. Segmentations are used in Biomediq's own pipeline for further analysis. This includes atrophy calculations between time points and analysis of shape and texture to distinguish AD from other clinical groups, and ultimately test if AD medication is effective. Texture and shape analysis are done from hippocampal segmentations, whereas atrophy calculations are done using hippocampal segmentations as well as other subcortical structures. FreeSurfer will be used as a reference method is this thesis, accordingly it is not the intension to give a detailed description of the steps. A part of the FreeSurfer pipeline will be used to preprocess the images, these steps are explained in Chapter 4.

Segmentation of subcortical structures are done using both cross-sectional and longitudinal FreeSurfer (v.5.1.0) [4]. In both methods, a neuroanatomical label is assigned to each image voxel. Longitudinal FreeSurfer uses information from more than one time point simultaneously to do segmentation of a single time point, whereas cross-sectional FreeSurfer does segmentation based on a single time point. The FreeSurfer pipeline contains 31 steps in total. The following main steps are performed in FreeSurfer:

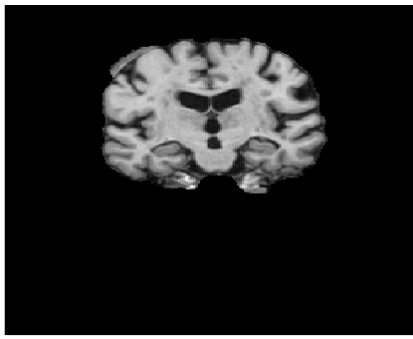
1. Affine transformation to a standard space (atlas).
2. Bias field correction.
3. Intensity normalization.
4. Skull stripping (whole brain segmentation).
5. Linear and non-linear registration to a brain atlas.
6. Final labeling of brain structures.

At Biomediq A/S, the entire FreeSurfer program package is run for every dataset. However, it is primarily the subcortical segmentations and the intensity images after bias field correction that are used to make further analysis. Figure 1.5 shows some images and segmentations of a single subject obtained using FreeSurfer. The corresponding hippocampal segmentations in 3D are illustrated in Figure 1.6.

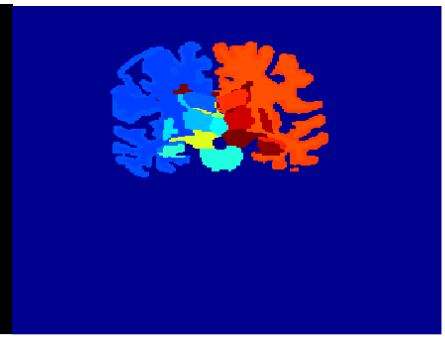


(a) Original.

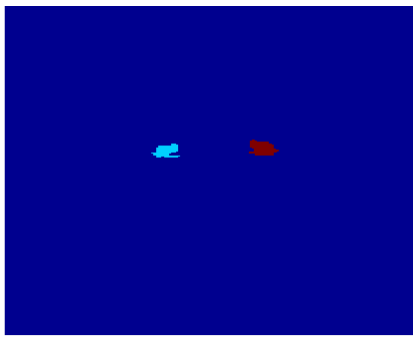
(b) Bias field corrected.



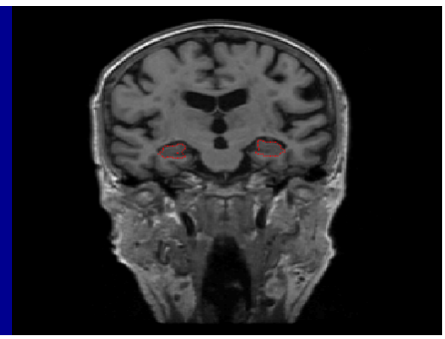
(c) Skull-stripped.



(d) Subcortical segmentations.



(e) Hippocampus.



(f) Hippocampus border superimposed on bias field corrected image.

Figure 1.5: Different images obtained using FreeSurfer from T1-weighted MRI. e) Left hippocampus: light blue. Right hippocampus: red.

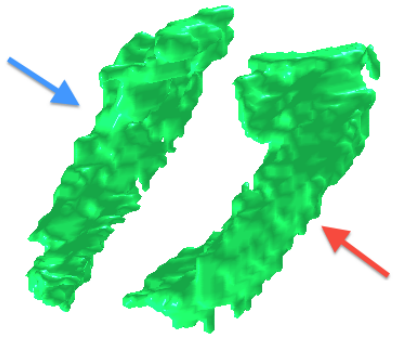


Figure 1.6: 3D hippocampus segmentation using FreeSurfer. Right: Red arrow. Left: Blue arrow. The same subject as illustrated in Figure 1.5.

1.2.1 Undesirable features

Overall FreeSurfer is reliable. However, to ensure satisfying segmentations it is necessary to visually inspect all subjects for segmentation errors. Below are listed some of the experienced problems and undesirable features regarding hippocampal segmentation.

1. Hippocampal segmentation is too rough in some slices, Figure 1.7. In some cases, this is due to bad image contrast. Generally, FreeSurfer has difficulties in segmenting brains of elderly subjects and especially AD brains due to pathological changes observed in these subjects, e.g. enlarged ventricles and whole brain atrophy, Figure 1.3.
2. Developed to segment all brain structures, which potentially hampers a good segmentation of a specific structure (hippocampus).
3. Computation duration takes 11+ hours.
4. Limited access to change parameter settings and no possibility to change source code.
5. Original image resolution is conformed to 1x1x1 mm^3.
6. Voxels are interpolated during registrations and intensities are changed during e.g. bias correction and intensity normalization, which affects especially texture analysis.

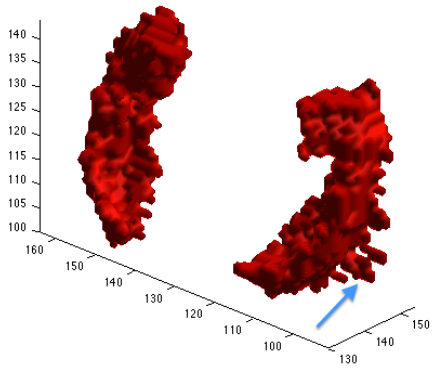


Figure 1.7: 3D illustration of a segmentation from FreeSurfer. The segmentation of hippocampus is too rough (blue arrow).

1.3 Project goals

Biomediq's goal is to have their own segmentation pipeline, accordingly, they aim at eliminating the use of FreeSurfer. A segmentation pipeline includes preprocessing as well as segmentation. In this project, focus will be on segmentation. Based on the company needs, the project goals are:

1. Robust automated segmentation of T1-weighted MRI subcortical brain structures.
2. Main focus in segmentation of hippocampus, but the method should potentially be extended to other structures if needed. It is better to improve the segmentation of hippocampus significantly, than making a mediocre segmentation of all structures.
3. Use two state-of-the-art methods and compare to FreeSurfer.
4. Computation duration preferably faster than 11+ hours.
5. More control with segmentation process. Capability to change parameters and code.

Hippocampal segmentation has been the focus of this thesis, accordingly other subcortical structures have not been segmented.

1.4 Thesis overview

The following gives a brief overview of the chapters and appendices in the thesis.

- **Chapter 2 - State-of-the-art** summarizes the current state-of-the-art segmentation methods and their performance. This leads to a selection of two methods.
- **Chapter 3 - Data and Atlases** introduces the data and the atlas used for segmentation.
- **Chapter 4 - Preprocessing** describes the MRI preprocessing (biascorrection and skull-stripping) and the transfer of atlas labels and MRI to different segmentation spaces using affine and rigid registrations.
- **Chapter 5 - Segmentation** covers theory of the two methods used to segment hippocampus.
- **Chapter 6 - Parameter and method selection** estimates the optimal method parameters based on leave-one-out cross-validation with two atlas sets. Based on this analysis an evaluation is made and one method is selected to segment the entire dataset.
- **Chapter 7 - Final results** evaluates the segmentation results based on volume and atrophy by making a comparison to FreeSurfer segmentations. A statistical analysis is performed. Finally, the results are discussed.
- **Chapter 8 - Conclusion** gives the conclusion together with a proposal for future work.
- **Appendix A** contains tables with demographics of the atlases used.
- **Appendix B** contains tables and figures of the statistical analysis in Chapter 7.
- **Appendix C** contains a CD with the volume segmentations at several time points and the atrophy scores between time points. Furthermore, the Non-Local Patch-based segmentation source code is included. The CD also contains the R-code and the m-code made for statistical analysis.

Chapter 2

State-of-the-art

Established methods for segmenting brain volumes from MRI can be classified into two groups: Basic tissue classification and anatomical segmentation, Figure 2.3, row one.
Automated basic tissue classification is done based on intensity information and can be used to distinguish brain from non-brain, and within the brain, White Matter (WM), Grey Matter (GM) and CerebroSpinal Fluid (CSF) [14].

Automated segmentation of subcortical brain structures is comparatively challenging. Signal intensities alone are not sufficient to distinguish between structures, because they show considerable overlap, [4], [31]. Even distinct anatomical structures can have the same MRI signal properties. Figure 2.1 illustrates the intensity histograms of different brain structures from T1-weighted MRI. The overlap of hippocampus (Hp) and and the structure lying next to it, amygdala (Am), is almost total, and many of the other structures are considerable overlapping. Furthermore, a structure can be composed of more than one tissue type, which prevents the use of simple intensity based approaches. Hippocampus is especially difficult to segment due to its small size, high variability, low contrast and discontinuous boundaries on MRI [8]. The hippocampal surface volume accounts for approximately 10 % of the volume of the entire structure. Therefore, even small impressions in segmentation can affect the result significantly.

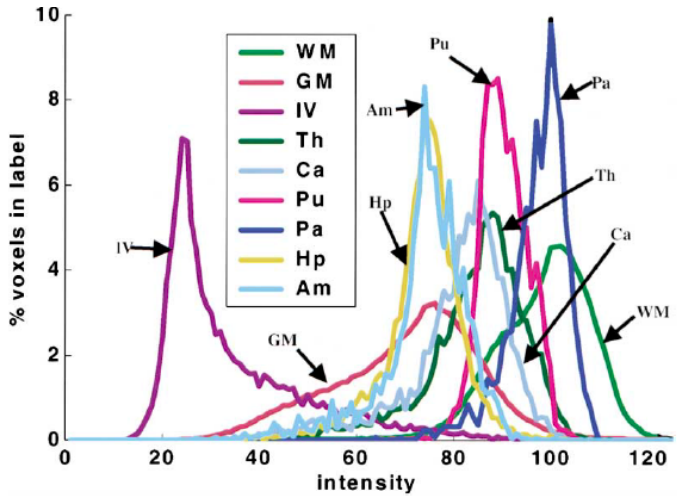


Figure 2.1: Intensity histograms from T1-weighted MRI for White Matter (WM), cortical Gray Matter (GM), Lateral Ventricle (IV), Thalamus (Th), Caudate (Ca), Putamen (Pu), Pallidum (Pa), Hippocampus (Hp) and Amygdala (Am) [4].

Figure 2.2 shows the MRI of both amygdala and hippocampus in a slice, together with segmentations which distinguishes between the two structures. The images illustrate the difficulty in distinguishing between the structures - not all edge structures are visible on MRI, e.g a part of hippocampus' border with amygdala is usually invisible.

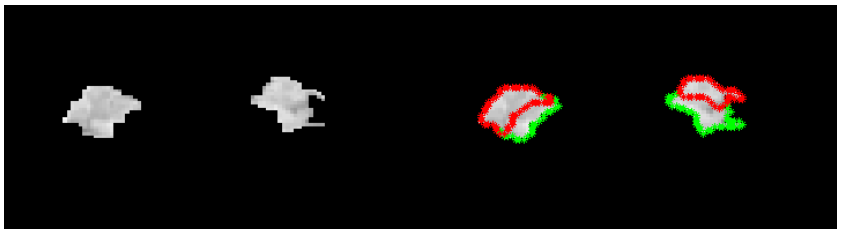


Figure 2.2: MRI slice, coronal view. Left: MRI of amygdala and hippocampus. Right: corresponding segmentation. Red: Amygdala. Green: Hippocampus.

Automatic 3D subcortical methods can incorporate the use of statistical models of intensity and shape, machine learning techniques, level sets, region growing or anatomical atlases. Most techniques can be divided into 3 categories. 1) Deformable models. 2) Appearance-based models or 3) Atlas-based/template-warping techniques, Figure 2.3, row two.

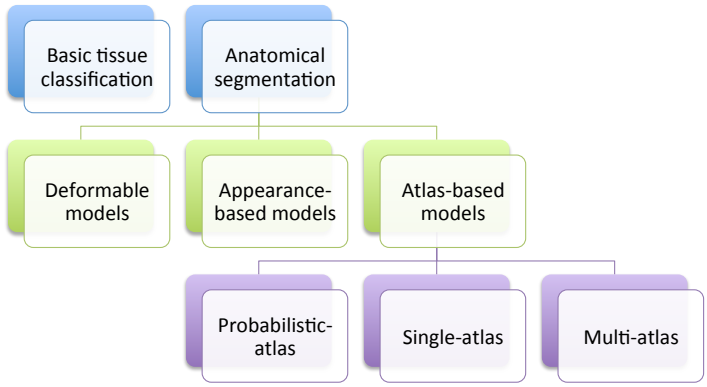


Figure 2.3: Overview of different classification methods used to automatically segment brains.

Deformable models are curves or surfaces in an image domain, which can move within the influence of different forces (from the model itself or from the image). In [13] a deformable contour technique is used to customize a balloon model to a subject's hippocampus.

Appearance-based models establish correspondences across a training set and learns the statistics of shape and intensity variations using PCA models [5].

In atlas-based segmentation, prior knowledge is available in an atlas. An atlas is a manual annotation of anatomical structures of interest by expert operators, accordingly additional information is augmented besides the voxel intensities alone. An atlas MRI corresponds of two images: MRI and the corresponding manual annotations/labels. Different forms of atlases can be used for segmentation, 1) a probabilistic atlas 2) a single-atlas or 3) multi-atlases, Figure 2.3, row three.

Probabilistic atlases contain pre-computed statistics of a set of labeled images, atlases, which are registered using non-rigid registration. In probabilistic atlases the cross-subject averaging may remove potentially useful information. The

probabilistic atlases can be used to incorporate structure specific models using Markov Random Fields in a Baysian framework [4].

In single- and multi-atlas techniques, the atlas MRI (training image) is registered to a test image (image to be segmented) usually by optimizing an intensity-based similarity measure. The transformation is then used to deform the atlas labels to the test image. However, the segmentation result using one atlas is sensitive to the manual segmentation, the image registration procedure and considerable differences between the test image and the atlas image anatomy [1]. One manual labeling is seldom enough to make a rich representation of an entire population. Figure 2.4 shows some examples from the ADNI dataset (explained in Chapter 3) illustrating a wide range of morphological variations in hippocampus. Preferable, these variations should all be represented in the atlas used.

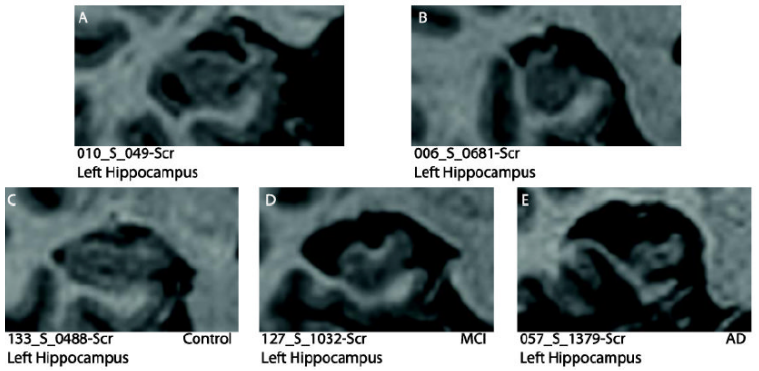


Figure 2.4: Examples from the ADNI dataset (explained in Chapter 3) which illustrates the wide range of morphological variation in hippocampi. A) A large hippocampal cyst and lack of temporal horn. B) Malrotation (tall and narrow). C) Normal hippocampus. D) MCI hippocampus (considerable atrophy) and E) AD hippocampus (atrophy) [25].

To account for the anatomical variations between subjects, the segmentation can be improved by using a multi-atlas segmentation approach, where multiple atlases are registered to the test image and the deformed labels are combined by label fusion strategies. The steps in a typical multi-atlas approach can be seen in Figure 2.5. Multi-atlas segmentation is reported to be among the best when dividing the whole brain into multiple segments [14] or when targeting individual structures, e.g hippocampus [31], [6]. Multi-atlas segmentation has shown to outperform other state-of-the-art methods [5].

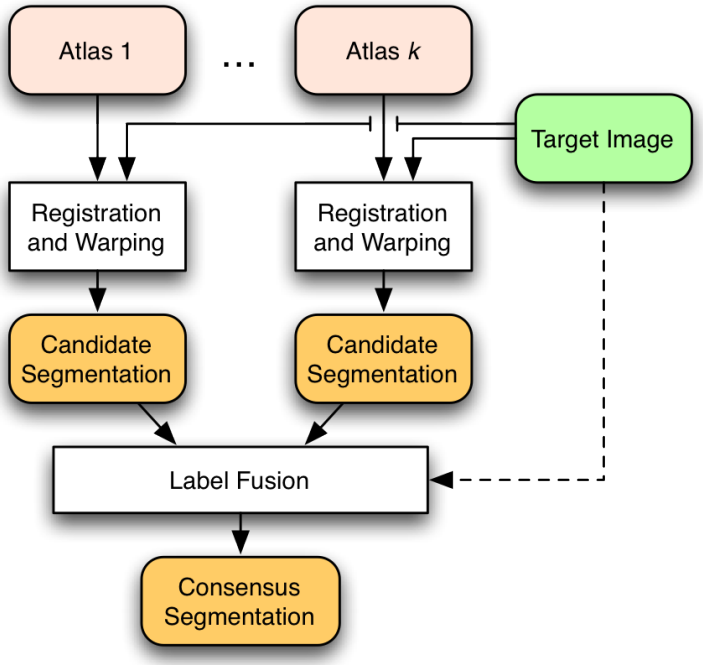


Figure 2.5: Steps in a typical multi-atlas segmentation method with label fusion [18].

In most multi-atlas methods the registration is non-rigid, which means the computational cost of registering many atlas images to a test image is high. Furthermore, segmentation based on dissimilar images can lead to incorrect segmentation based on the choice of label fusion strategy. Therefore, Aljabar et al. [1] proposed a method using only the most similar atlases. The similarity measure was either based on image similarity measures prior to detailed non-linear registration or based on meta-data such as subject age. In [33] a low dimensional representation of the data is used to find morphologically similar datasets. An image is only registered to similar atlases, and label propagation is performed, creating new segmentations which can serve as atlases in further registrations and label propagations.

Different label fusion strategies exist. The simplest fusion technique is *Majority Voting*. Each voxel in the test image are given the label that is represented most times in the warped atlases. In weighted averaging the training subjects more similar to the test subjects carry more weight in the final label fusion. The similarity measure includes using the entire image to determine one global weight for each training subject, employing local image intensities to determine the weight of each voxel or combining the segmentations based on a probabilistic model e.g. STAPLE [28].

Recently, Non-Local Patch-based segmentation techniques have been proposed. These models do not need the computational heavy non-rigid registrations. A label is obtained for every voxel by using similar image patches from coarsely aligned atlases using affine registrations [8].

2.1 Method choice

Since multi-atlas techniques have outperformed other state-of-the art methods, multi-atlases techniques will be used in this thesis. Registration is often computational heavy in these methods. If a method should be used in the clinic, the segmentations should preferably be available immediately after the images were acquired. Therefore, it will be analyzed how a less computational heavy method only using affine registrations to align images performs, compared to a method using non-rigid registrations. The method using affine registration will be an implementation of the Non-Local Patch-based segmentation from [8]. Of non-rigid methods, BrainFuseLab [28] has shown promising results and is furthermore developed to use FreeSurfer preprocessed images as input. The BrainFuseLab code is available online [17], whereas a Non-Local Patch-based method must be implemented. To make a fair comparison, both methods should use the same atlases. The methods will be explained in Chapter 5.

Chapter 3

Data and Atlases

3.1 Data

Data used in the preparation of this thesis were obtained from the Alzheimer's disease Neuroimaging Initiative ADNI database (adni.loni.usc.edu). The 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, biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early 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.

Three different large ADNI studies have been conducted - *ADNI-1*, *ADNI-2* and *ADNI-go*. Three diagnostic groups are available in the ADNI data - People with Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI) and Cognitively Normal (CN). The pathological differences between these groups are explained in Section 1.1. The ADNI data include clinical, imaging, genetic and biochemical biomarkers. In this thesis, only T1-weighted 1.5T MRI will be analyzed. Since

the ADNI study is a multisite study, the T1-weighted MRIs are acquired at different MRI systems (General Electric (GE) Healthcare, Philips Medical Systems, Siemens Medical Solutions) with a repeated Magnetization Prepared Rapid Gradient Echo (MP-RAGE REPEAT) sequence. The image dimensions vary from scanner to scanner with resolution in the range $[0.94, 1.35]\times[0.94, 1.35]\times 1.2mm^3$.

A standardized part of the ADNI-1 dataset is used, *2-year annual complete*, (baseline, month 12 and month 24 scans). A standardized dataset is made to ensure a meaningful methodological comparison, thereby mitigating the risk that differences in algorithm performance are an artifact of the use of different input [34]. The dataset consist of 504 subjects, 169 CN, 234 MCI and 101 AD and will thus be denoted *ADNI504*. All subjects were included in the standardized dataset if the MRI of at least one of the two replicate T1-weighted scans passed the QC control. Each subject should have all their scans performed at the same scanner, due to variations in images not only from system to system, but also from scanner to scanner. The mean age, gender and Mini-Mental State Examination (MMSE) score of the subjects in the three diagnostic groups at baseline are listed in Table 3.1. MMSE is a cognitive test, including questions in arithmetic, memory and orientation, used to screen for cognitive impairment and to follow cognitive changes in a person over time. It is possible to achieve a maximum MMSE score of 30 points. Table 3.1 includes basic statistics between groups. It should be noted that the MCI group contains a significantly larger percentage of men than the CN group and the AD group, respectively.

	Group		
	CN(n=169)	MCI(n=234)	AD(n=101)
Age, yr $\pm\sigma$	76.0 ± 5.1	74.9 ± 7.0	75.3 ± 7.4
Men (%)	50.9	66.7	50.5
MMSE $\pm\sigma$	29.2 ± 1.0	27.1 ± 1.7	23.2 ± 1.9
	Statistics (p-value)		
	CN vs. MCI	CN vs. AD	MCI vs. AD
Age, yr $\pm\sigma$	0.066	0.318	0.631
Men (%)	0.002	1	0.008
MMSE $\pm\sigma$	<0.001	<0.001	<0.001

Table 3.1: ADNI504: Baseline demographics (age, gender) and clinical parameters (MMSE) as well as statistics between groups. χ^2-test was applied to obtain the p-value for gender while two sample two sided t-tests were used for the remaining parameters.

3.2 Atlases

To get good segmentation results it is important to select an atlas dataset which represents the variability that corresponds to the population to be segmented. Not many atlases are available for download, and the few available are most often based on healthy young subjects, who have brains dissimilar to the population of greatest risk developing AD, elderly people.
It is hard to distinguish hippocampus from its surrounding structures, even experts do not agree on an unequivocal definition. Therefore, it is extremely difficult to establish ground truth by manual segmentations which is reflected in various definitions of atlases used for automated segmentation. An atlas set consists of two sets of images: 1) Manual labels and 2) the corresponding MRI.

3.2.1 Harmonized Hippocampal Protocol

A new initiative, *A Harmonized Protocol for Hippocampal Volumetry: an EADC-ADNI Effort* [12], has been establish in recent years to make a streamlined manual segmentation protocol. The goal is to agree on the anatomical landmarks and measurement procedure. By elaborating this protocol, it will be possible to directly compare the effect of different drugs in slowing down neurodegenerative processes and further define the golden standard for automated segmentations [22].

A web-based qualification system is made, which allows tracers worldwide to learn manual hippocampal segmentation based on the harmonized protocol. In connection with the protocol, manual segmentations of at the moment 100 ADNI images (35 more to come) have been released. The released labels cover a wide range of physiological variability and are therefore suited for training and validation of automated algorithms.

A subset of these manual annotations will serve as the atlas set in this thesis. These manual segmentations are chosen as atlas set in this work, because they include both AD, MCI and CN of elderly subjects and they are as close as one can get to a hippocampal segmentation golden standard. Since the labels have just been made publicly available in August 2013, this work will be one of the initial studies that evaluates how the labels perform as atlas set in state-of-the-art automated segmentation methods. The manual segmentations will be used as atlas set in the two methods explained in Chapter 5. A coronal, sagittal and transversal view of a CN, MCI and AD subject is shown in Figure 3.1. The manual hippocampus labels (red) are superimposed on the underlying MRI. The corresponding 3D illustrations of the manual labels can be seen in Figure 3.2.

Two different atlas sets are used in this thesis. Both sets are subsets of the released manual labels from the Harmonized Hippocampal Protocol (HHP). *Atlas15* includes 15 manual segmentations - these scans are not part of ADNI504. *Atlas40* includes 40 manual segmentations, some of them are part of ADNI504. Information of each subject in these atlas sets can be found in Appendix A. The mean age, gender, MMSE score and hippocampal volume of the cognitive state (CN, MCI and AD) for the two atlas sets can be seen in Tables 3.2 and 3.3.

	Group		
	CN(n=6)	MCI(n=2)	AD(n=7)
Age, yr $\pm\sigma$	76.3 ± 7.9	72.7 ± 1.1	75.7 ± 8.0
Men (%)	33.4	100	77.8
MMSE $\pm\sigma$	28.7 ± 1.0	27.0 ± 1.4	24.3 ± 2.8
Volume $(mm^3) \pm \sigma$	8127± 1240	7953 ± 1392	6952 ± 772

Table 3.2: Atlas15: Age, gender, MMSE score and hippocampal size for CN, MCI and AD.

	Group		
	CN(n=12)	MCI(n=11)	AD(n=17)
Age, yr $\pm\sigma$	76.9 ± 6.2	70.9 ± 6.8	74.2± 8.6
Men (%)	41.7	54.6	47.1
MMSE $\pm\sigma$	28.8 ± 1.2	27.6 ± 1.2	24.0 ± 2.7
Volume $(mm^3) \pm \sigma$	8176 ± 996	7708 ± 769	6887 ± 1080

Table 3.3: Atlas40: Age, gender, MMSE score and hippocampal size for CN, MCI and AD.

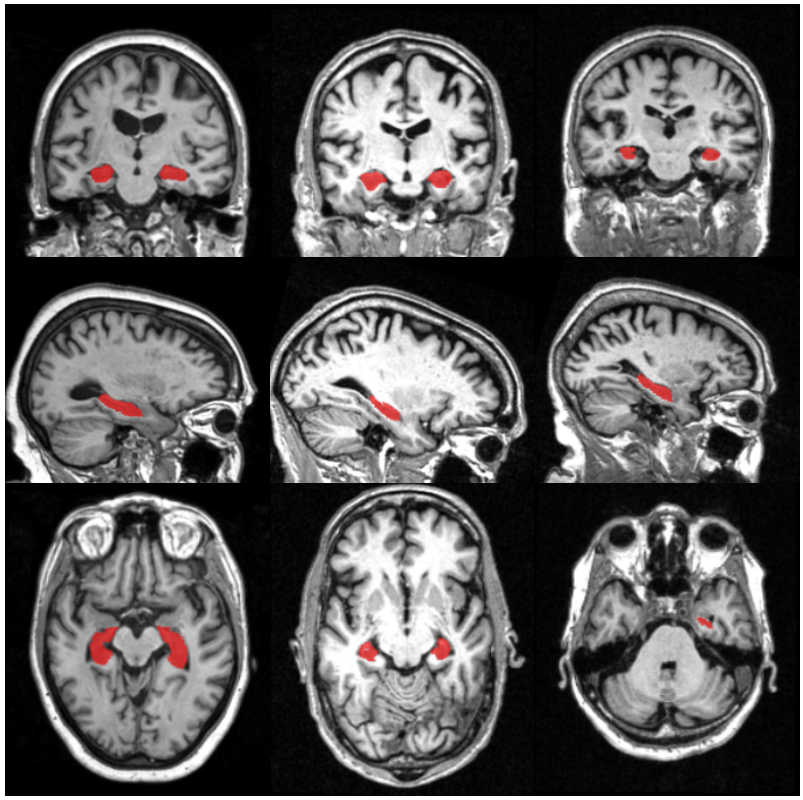
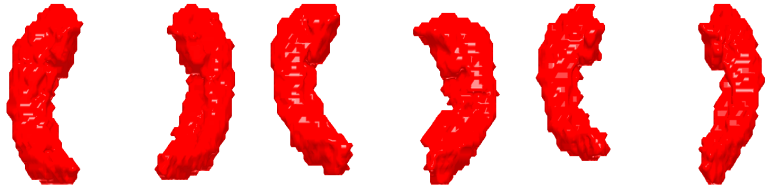


Figure 3.1: Coronal, sagittal and transversal view of manual labels from the Harmonized Hippocampal Protocol. The atlas set consists of manual labels of hippocampus (red) and the underlying MRI. CN: Column 1. MCI: Column 2. AD: Column 3. View (X=70,Y=117, Z=69).



(a) CN (b) MCI (c) AD

Figure 3.2: 3D illustrations of the manual labels from Figure 3.1.

3.2.2 FreeSurfer atlas

Cross-sectional and longitudinal FreeSurfer will be used as reference methods in this thesis. Both FreeSurfer methods uses the same probabilistic atlas, Chapter 2. Therefore, it has not been possible to include Harmonized Hippocampal Protocol atlases in this segmentation method, thus the hippocampal definition in the FreeSurfer atlas is different than the atlas set used in BrainFuseLab and Non-Local Patch-based segmentation. 39 subjects are used to build the FreeSurfer atlas. They are a combination of healthy subjects as well as patients of various ages with probable or questionable AD [4]. The atlas includes 37 subcortical brain structures, and segmentations of all 37 structures are thus available. In this thesis, only hippocampal segmentations will be considered and serve as a reference.

Chapter 4

Preprocessing

Due to large intensity differences in MRI, the test data and the training data (atlases) must be preprocessed before segmentation is carried out with the various methods used in this thesis. Initially, the atlases and the preprocessed MRIs are not in the same space, thus they must be transformed to a common space prior to segmentation. Preprocessing will be explained in this chapter and involves:

1. MRI preprocessing (*bias field correction* and *skull-stripping*).
2. Transformation of atlas labels and preprocessed MRI to a common segmentation space.

4.1 MRI preprocessing

MRI preprocessing is done with FreeSurfer (v.5.1.0). Cross-sectional and longitudinal FreeSurfer segmentations are thus obtained from the same preprocessed images as the segmentations obtained with the two methods explained in Chapter 5. This reduces the factors which can explain differences in segmentation results.

The first step is to conform the original MRI resolution, which is in the range [0.94, 1.35] × [0.94, 1.35] ×1.2 mm^3, to isotropic voxels, 1 × 1 × 1 mm^3. The image dimensions are changed to 256×256×256 voxels. During preprocessing, intensity normalization is done multiple times, where the MRI is scaled according to peak values within White Matter (WM), Grey Matter (GM) and CerebroSpinal Fluid (CSF).

4.1.1 Bias Field Correction

A MRI varies in both intensity and contrast across the 3D image. This spatial intensity inhomogeniety is called the *bias field effect*. The bias field effect is proportional to the scanners field strength and is caused by the Radio Frequency field inhomogeneities. Due to the bias fields effect, intra-class homogeneity can not be assumed and accordingly identical tissue types will vary in intensities as a function of their spatial location. This is an undesirable condition for any segmentation method, where intensity information is used to classify voxels into different tissue types. The bias field effect is unique for each subject, which makes it challenging to correct it. FreeSurfer uses the non-parametric non-uniform intensity normalization, N3 [30], to correct for the bias field effect. The method is based on the following assumed model of MRI formation:

$$v(x) = u(x)f(x) + n(x) \tag{4.1}$$

Where x is the location, v is the measured signal, u is the true signal, f is an unknown smoothly varying bias field, and n is white gaussian noise. To correct for the bias field, f must be estimated. In Equation 4.1 the bias field is interfered by both an additive and multiplicative component, therefore, a noise-free additive model is used instead, with the notation $\hat{u}(x) = log(u(x))$:

$$\hat{v}(x) = \hat{u}(x) + \hat{f}(x) \tag{4.2}$$

U,V and F are the probability densities of $\hat{u}$, $\hat{v}$ and $\hat{f}$, respectively. $\hat{u}$ and $\hat{f}$ are approximated uncorrelated random variables, and the distribution of their sum is found by convolution:

$$V(\hat{v}) = F(\hat{v}) * U(\hat{v}) = \int F(\hat{v} - \hat{u})U(\hat{u})d\hat{u} \tag{4.3}$$

The task is to restore the frequency content of U, to get from the observed

distribution V to the true distribution U. However, it is unknown which frequency components of U that need to be restored. The approach is to find the smooth slowly varying field, $\hat{f}$, that maximizes the frequency content of U. This is done by sharpening the distribution of V, estimate the corresponding $\hat{f}$, which produces a distribution of U close to the one suggested. F is assumed to be Gaussian, having zero mean and a given variance, which means it is only necessary to search the space of distributions U corresponding to the properties of F. A MRI prior to and after bias field correction can be seen in Figure 1.5 (a) and (b), respectively.

4.1.2 Skull-stripping

Whole-brain segmentation (skull-stripping) is an important discipline in analysis of neuroimaging data. During skull-stripping brain tissue is removed from non-brain tissue such as skull, eyeballs and skin. In FreeSurfer a watershed algorithm combined with a deformable model is used to peel the skull [29]. Two assumptions are made:

1. Connectivity of WM is assumed, bordered by GM and CSF.
2. The brain surface, which distinguish non-brain from brain, is a smooth manifold with relatively low curvature.

Before the watershed algorithm is applied, some parameters must be computed, these include an upper intensity bound for CSF, the centroid of the brain, an average brain radius, lower and upper bound for white matter intensity and a global brain minimum within a cubic region centered at the centroid of the brain.

The watershed algorithm:
Because white matter connectivity is assumed, WM surrounded by lower intensity GM and even lower intensity CSF in T1-weighted MRI, can be interpreted as a hill in a 3-dimensional landscape. By inverting the grey-scale values, the WM hill becomes a valley. The concept of pre-floating height is introduced. Prior to finding a connectivity path, each basin in the landscape is flooded up to a certain height above its bottom, the pre-floating height h_{pf}. The default value of h_{pf} is 25, corresponding to 25 % of the maximum intensity. If the pre-floating height is at a higher altitude than the the basin border, the basin cannot hold water, and it will be merged with the deepest neighboring basin. If it holds water, it will be regarded as a separate region. Voxels are connected in a path, even if a lower intensity than the darkest of the two points are present up

to a maximum difference, the pre-floating height. After the watershed transformation has been applied, the segmented volumes still contain non-brain tissue such as CSF, some parts of the skull and often the entire brain-stem. The result of the watershed algorithm can be seen in Figure 4.1. The output of the watershed algorithm will serve as an initialization for a deformable model.

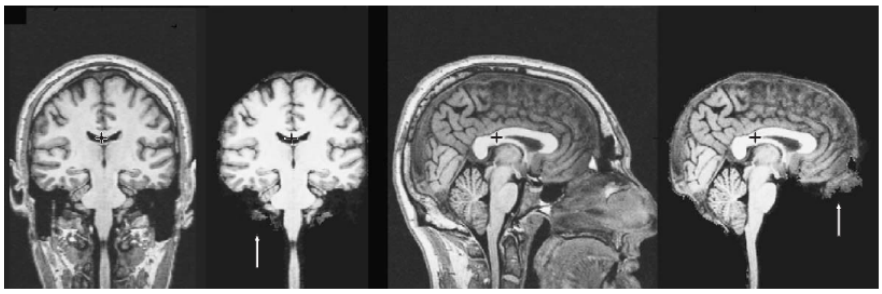


Figure 4.1: Before and after watershed transformation. The black cross points out the centroid of the brain. White arrows indicate non-brain regions, which have not been removed by the watershed transformation [29].

Deformable surface algorithm:
The segmented volume from the watershed algorithm is used to initialize a deformable balloon-like template using active contours. The initial brain surface model is an icosahedron with 10242 vertices. This template is centered at the centroid of the brain, with a radius that includes the whole previously segmented brain. The template is then gradually transformed through a series of iterative steps. In each iteration the coordinate of each vertex is updated according to three forces, a smoothness force F_s, a MRI-based force F_{MRI} that drive the template towards the true brain boundary and an atlas force F_A that ensures the deformed template has the shape of a brain within a certain tolerance. An example of a deformation process can be seen in Figure 4.2. The deformed template is then used to skull-strip the three dimensional MRI by removing the voxels outside the estimated surface, 4.3.

Figure 4.2: Template deformation process. Left: initial template (icosahedron). Right: Final template [9].

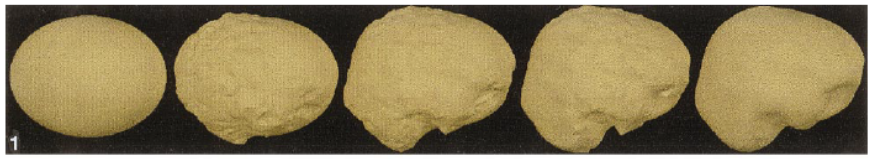
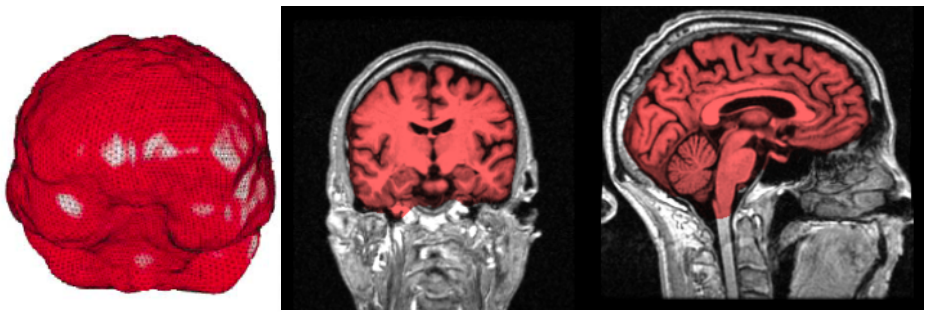


Figure 4.3: Final skull-stripping: Final deformed template (left) [29]. Middle and Right: Skull-stripped brain (red) superimposed on the original MRI



4.2 Registration and label transformation

The harmonized hippocampal standard manual segmentations are done by expert operators in a standard space called MNI space. Prior to the manual segmentation, the MRIs have been aligned to a template containing an average of 152 brains in MNI space. This template can be seen in Figure 4.4

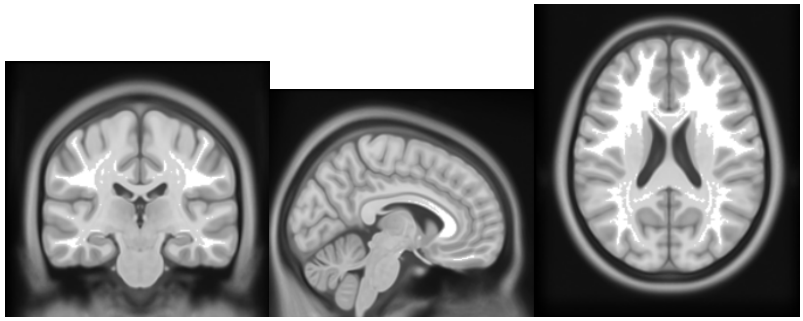


Figure 4.4: Coronal, sagittal and transversal view of the MNI template found by averaging of 152 brains. Originally, the HHP atlases are in this space. Image dimensions: 197×233×189.

To use the manual segmentations as atlases in the automated segmentation methods, it is necessary to get them and the FreeSurfer preprocessed MRIs to a common space. Since the preprocessed MRI is already in FreeSurfer space, the labels will be taken to this space. This involves transforming labels from image dimensions 197×233×189 to 256×256×256. The atlases will in this thesis be aligned using two different registrations. The first is a *rigid-body* transformation, whereas the other is an *affine* transformation.

Two steps are involved in registering a pair of images, *registration* and *transformation*. In the registration, a set of parameters describing the transformation are estimated. In the transformation, one of the images is transformed according to the estimated parameters. Both the registration and transformation are done using SPM [21],[2].

4.2.1 Rigid-Body Registration

Rigid-body or rigid transformations are a subclass of affine transformations. For each point in an image (x_1, x_2, x_3) an affine mapping into the co-ordinates of another space (y_1, y_2, y_3), can be represented as:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ 1 \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{bmatrix} \tag{4.4}$$

Rigid-body registrations consist of only translation and rotation and involves estimating 6 parameters (3 for translation, 3 for rotation).

Translation: The translation of a point $\mathbf{x}$ by $\mathbf{q}$ units, is given by:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & q_1 \\ 0 & 1 & 0 & q_2 \\ 0 & 0 & 1 & q_3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{bmatrix} \tag{4.5}$$

Rotation: The object can be rotated around three orthogonal planes (axes) in three dimensional images. Rotation matrices of q_4, q_5 and q_6 radians around the x-axis, y-axis and z-axis respectively are given by:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & cos(q_4) & sin(q_4) & 0 \\ 0 & -sin(q_4) & cos(q_4) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} cos(q_5) & 0 & sin(q_5) & 0 \\ 0 & 1 & 0 & 0 \\ -sin(q_5) & 0 & cos(q_5) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} and \begin{bmatrix} cos(q_6) & sin(q_6) & 0 & 0 \\ -sin(q_6) & cos(q_6) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{4.6}$$

Multiplication of these matrices combines rotations. The order of the multiplication influences the result.

4.2.2 Affine Registration

In affine transformation 12 parameters have to be estimated (3 translation, 3 rotation, 3 scaling and 3 shearing). The translation and rotation is calculated in the same way as described for rigid-body registration.

Scaling: Scaling is needed to change the size of the image. Scaling can be represented as:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ 1 \end{bmatrix} = \begin{bmatrix} q_7 & 0 & 0 & 0 \\ 0 & q_8 & 0 & 0 \\ 0 & 0 & q_9 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{bmatrix} \tag{4.7}$$

Shear: Shear mapping by parameters q_{10}, q_{11} and q_{12} are given by:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & q_{10} & q_{11} & 0 \\ 0 & 1 & q_{12} & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{bmatrix} \tag{4.8}$$

4.2.3 Optimization

Optimization is done to find the optimal parameters $\mathbf{q}$. One image (source image) is spatially transformed so it matches another image (reference image) by minimizing or maximizing some function or parameter. The usual approach is to do iteratively searching from an initial parameter estimate. At each iteration a judgement is made, before moving on to the next iteration. In both the affine and rigid-body registrations, Gauss-Newton optimization is done based on minimizing the *sum of squared differences* (SSD) dissimilarity measure.

The Gauss-Newton idea consists of linearizing the function (by Taylor expansion to first order). The parameters are updated by solving a set of linear equations obtained from setting the first order derivatives equal to zero.
$b_i(\mathbf{q})$ is the SSD describing the difference between the source and the reference image at voxel i, when the model parameters have values $\mathbf{q}$. The method estimates the values of $\mathbf{t}$ in order to minimize $\sum_i b_i(\mathbf{q}-\mathbf{t})^2$. This is done from the following sets of equations:

$$\begin{bmatrix} \frac{\partial b_1(\mathbf{q})}{\partial q_1} & \frac{\partial b_1(\mathbf{q})}{\partial q_2} & . & . \\ \frac{\partial b_2(\mathbf{q})}{\partial q_1} & \frac{\partial b_2(\mathbf{q})}{\partial q_2} & . & . \\ . & . & . & . \\ . & . & . & . \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \\ . \\ . \end{bmatrix} \simeq \begin{bmatrix} b_1(\mathbf{q}) \\ b_2(\mathbf{q}) \\ . \\ . \end{bmatrix} \tag{4.9}$$

The parameters $\mathbf{q}$ are updated using an iterative scheme. For iteration n the parameters $\mathbf{q}$ are updated as:

$$\mathbf{q}^{(n+1)} = \mathbf{q}^{(n)} - (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b} \tag{4.10}$$

where

$$\mathbf{A} = \begin{bmatrix} \frac{\partial b_1(\mathbf{q})}{\partial q_1} & \frac{\partial b_1(\mathbf{q})}{\partial q_2} & . & . \\ \frac{\partial b_2(\mathbf{q})}{\partial q_1} & \frac{\partial b_2(\mathbf{q})}{\partial q_2} & . & . \\ . & . & . & . \\ . & . & . & . \end{bmatrix}, \mathbf{b} = \begin{bmatrix} b_1(\mathbf{q}) \\ b_2(\mathbf{q}) \\ . \\ . \end{bmatrix} \tag{4.11}$$

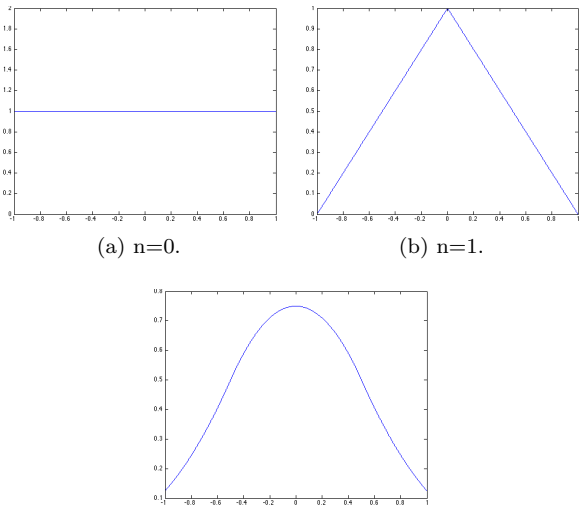
This is repeated until SSD can no longer be decreased or a maximum of 64 iterations is reached. However, the algorithm can be caught in a local minimum and therefore, there is no overall guarantee that the best global minimum is calculated.

4.2.4 Transformation

Interpolation for each voxel in a transformed image is used to determine the corresponding intensity in the original image. In this thesis, labels and images are interpolated using B-splines. B-splines are given by:

$$\beta^n(x) = \sum_{j=0}^{n} \frac{(-1)^j(n+1)}{(n+1-j)!j!} max\Big(\frac{n+1}{2} + x - j, 0\Big)^n \tag{4.12}$$

The degree, **n**, can be varied. **n**=0 corresponds to nearest neighbor interpolation, **n**=1 corresponds to trilinear interpolation and **n**=2 corresponds to cubic interpolation. In nearest neighbor interpolation the original voxel intensities are preserved. The value at each sample point is found by taking the value of the closest voxel. Trilinear and cubic interpolation is slower than nearest neighbor and uses the known intensities around the sample point to estimate the intensity at the sample point. The B-splines for $\mathbf{n} = \{0, 1, 2\}$ in 1-dimension is illustrated in Figure 4.5.



(a) n=0. (b) n=1.

(c) n=2.

Figure 4.5: B-splines with different degrees of **n** found using Equation 4.12. n=0: Nearest neighbor. n=1: Trilinear. n=2: Cubic.

The binary manual segmentations are transformed using nearest neighbor interpolation, whereas the MRI is transformed using cubic interpolation. Trilinear interpolation of the manual labels could have been an option, but this had involved determining an appropriate threshold to separate the transformed labels into object and background.

4.2.5 Illustrations

The atlases consist of manual segmentations and their corresponding MRI in MNI space. The same subjects are downloaded and preprocessed in FreeSurfer and are then in subject FreeSurfer (FS) space. Thus, the MRI of the same subjects are in two spaces.
Transfer of manual segmentations to a common segmentation space involves a combination of an intra-subject and an inter-subject registration whereas the FreeSurfer preprocessed MRI is transformed using only the inter-subject registration. The intra-subject registrations are always rigid-body transformations, whereas the inter-subject registration can be either a rigid-body transformation or an affine transformation. The registrations and transformations of MRIs and the manual labels for subject 003_S_0931 are illustrated in Figures 4.6, 4.7 and 4.8. The intra-subject registration, Figure 4.6, is done to find the transforma-

tion from MNI space to the the same subject in FS space registering original MRI conformed to isotropic voxels. The inter-subject registration, Figure 4.7, is done to find the transformation from a subject in FS space to another subject (Atlas) in FS space registering preprocessed FreeSurfer Norm MRIs (bias corrected, skull-stripped and intensity normalized). Transformations T1 and T2, Figures 4.6 and 4.7, are combined in Figure 4.8 by:

$$T3 = T2(M/T1) \tag{4.13}$$

If the labels are transferred from MNI space to Atlas FS space, Figure 4.8 , then M is a transformation matrix that maps voxel coordinates from the isotropic MNI image to a space whose axes have parallel image axes, origin is at the center of the image and distances are measured in millimeters. M is given by:

$$M = \begin{bmatrix} 1 & 0 & 0 & -DIM1/2 \\ 0 & 1 & 0 & -DIM2/2 \\ 0 & 0 & 1 & -DIM3/2 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{4.14}$$

where DIM1, DIM2 and DIM3 are the image dimensions.

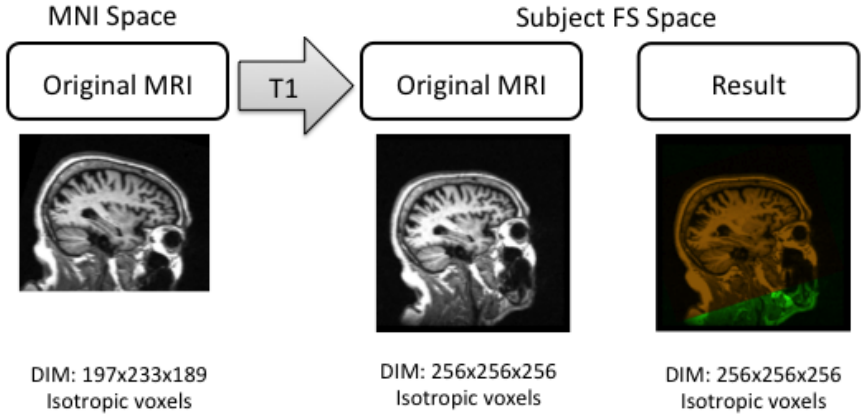


Figure 4.6: **Intra-subject Registration** using **rigid-body registration** (illustrated by T1 arrow) between two different spaces. Result: Red channel - Transformed MRI using T1 transformation and cubic interpolation. Green channel - Target image.

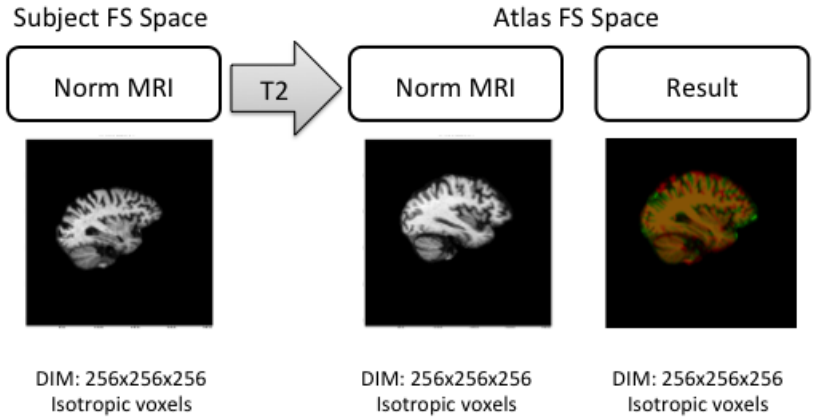


Figure 4.7: **Inter-subject Registration** using **affine registration** (illustrated by T2 arrow) between two different spaces. The registration can also be a rigid-body registration. Result: Red channel - Transformed Norm MRI using T2 transformation and cubic interpolation. Green channel - Target image (Norm MRI: bias corrected, intensity normalized and skull-stripped).

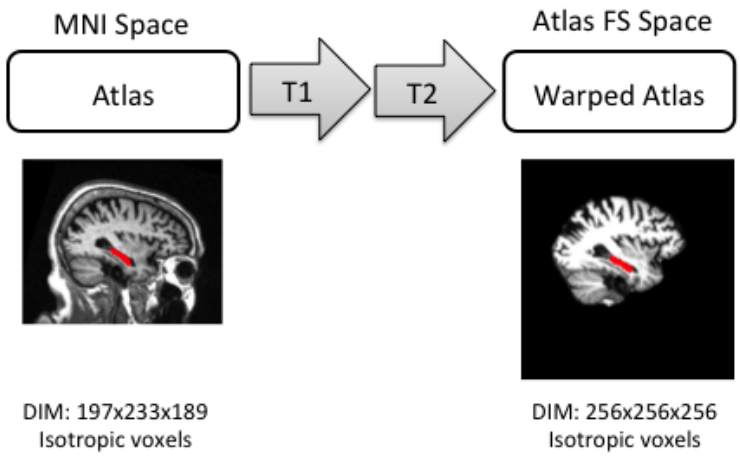


Figure 4.8: **Label transfer** combining T1 and T2 registrations from Figures 4.6 and Figures 4.7. Hippocampus (red) is superimposed on MRI in MNI space and on Norm MRI (bias corrected, intensity normalized and skull-stripped) in Atlas FS space.

Different combinations of the above illustrated transformations are used in the various segmentation methods in this thesis. However, in all cases, only one nearest neighbor interpolation is used to transfer the manual labels prior to segmentation to avoid loosing to much information. The transfer of labels for each segmentation method will be described in Chapter 5.

Since the binary labels are isotropic voxels in both their original space (MNI space) and FreeSurfer space after intra-subject transformation using only rigid transformation (T1), the number of voxels should ideally be the same in both spaces after nearest neighbor interpolation. In Table 4.1 the mean $\pm\sigma$ voxel difference (MNI space volume - FreeSurfer subject space volume) for the hippocampal labels can be seen for *Atlas15* and *Atlas40* introduced in Section 3.2. The table illustrates small volume differences, but they are within an acceptable range compared to the total hippocampal volumes of the two atlases, referenced in Table 3.2 and 3.3, typically 6500-8500 mm^3.

	Mean $\pm\sigma$
Atlas15	-1.47 ± 11.34
Atlas40	-0.03 ± 13.34

Table 4.1: Mean $\pm\sigma$ voxel difference (mm^3) of hippocampal labels (MNI space volume - volume of transformed labels to FS space) for *Atlas15* and *Atlas40*.

Chapter 5

Segmentation methods

Based on the content of Chapter 2, a multi-atlas Non-Local Patch-based segmentation method and a multi-atlas segmentation method using non-rigid registrations are tested. The Non-Local Patch-based segmentation (N-L Patch) is implemented from scratch and can be found on the CD in Appendix C, whereas the multi-atlas segmentation using non-rigid registration (BrainFuse-Lab) is available for download. The atlases from the Harmonized Hippocampal Protocol described in Chapter 3 will be used in both methods. Both methods use preprocessed images from FreeSurfer (bias field corrected, intensity normalized and skull-stripped) as explained in Chapter 4. This chapter describes the fundamental aspects of the segmentation methods.

5.1 Non-Local Patch-based segmentation

Segmentation is based on a Non-Local Patch-based framework using manual segmentations as priors [8] [7]. These models do not need the computational heavy non-rigid registrations, which are used in a majority of other multi-atlas approaches and are therefore considerably faster.

A label is obtained for every voxel by using similar image patches from coarsely aligned atlases. When the patch under study resembles a patch in an atlas,

their central voxels are considered to belong to the same structure. The patch that resembles the test patch is used in the estimation of the final label. Several patches from each atlas can be used during the label fusion of a single voxel, which increases the number of sample patches involved in the final label estimation compared to other multi-atlas approaches where each atlas typically weights a voxel ones. The term non-local indicates that the spatial distance between the center of the patches is not taken into account. The weight of each sample is solely depending on the intensity similarities between patches. The steps in the Non-Local Patch-based segmentation are explained below and can be seen in Figure 5.2. The optimal parameters will be found in Chapter 6.

Linear registration to one atlas:
Two sets of images need to be transformed to the same space, manual hippocampal labels and test MRI. The manual labels are transformed by combining T1 and T2, as illustrated in Figure 4.8, Chapter 4. Initially, the MRIs are preprocessed in FreeSurfer space and therefore only T2 transformation is needed to get the MRI to the atlas segmentation space, Figure 4.7. Both a rigid as well as an affine transformations are tried for T2, Chapter 6.

Initialization mask:
Due to computational issues, the segmentation will only be applied to voxels inside an initialization mask. The initialization masks are a union of the coarsely aligned atlases for left and right hippocampus, respectively.

Figure 5.1 illustrates three subjects registered with an affine registration to the same space with the initialization mask overlaid (blue).

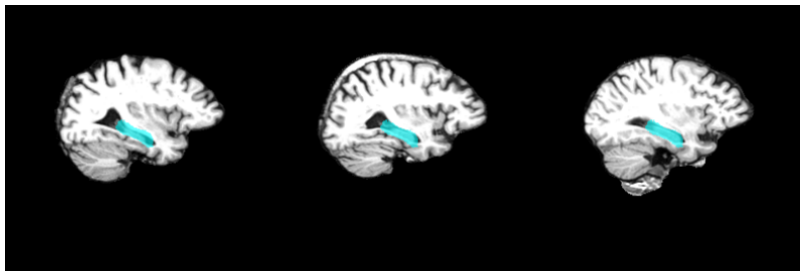


Figure 5.1: Three subjects registered to the same space with the initialization mask overlaid (blue).

Subject selection:
Due to computational cost, only a certain number of atlases, N, that resembles the subject under study the most, are used in the final non-local means label fusion. The similarity based measure used is the sum of squared differences

(SSD) across the initialization mask. SSD is used, because it is sensitive to e.g. contrast, which is an important factor in the label fusion. The subject selection is done for left and right hippocampus separately, which means that the same atlas subjects not necessarily contribute to both the right and the left hippocampus segmentation of a test subject.

Search volume:
A search for similar patches should be done in the entire image under study. However, this is computationally expensive. Therefore, only a limited search volume, V_i, is used defined as a cube centered at the voxel under study, x_i. Thus within the N closest selected atlases, the search for similar patches is in a cubic region around the voxel under study. The search volume must reflect the inter-subject variability, which can increase when pathological changes are present, e.g. in AD, and according to the structure under study.

Preselection:
In order to reduce the computational time, a preselection of patches are done discarding the most dissimilar patches. The preselection criteria is based on simple statistics such as mean and variance and can be seen below:

$$ss = \frac{2\mu_i\mu_{s,j}}{\mu_i^2 + \mu_{s,j}^2} \times \frac{2\sigma_i\sigma_{s,j}}{\sigma_i^2 + \sigma_{s,j}^2} \tag{5.1}$$

where μ represents the means and σ represents the standard deviation of patches centered on voxel x_i (voxel under study) and voxel $x_{s,j}$ at location j in subject s. If the value of ss is higher than a given threshold, the intensity distance between patches in the non-local means label fusion is calculated by Equation 5.3. The threshold is set to 0.95.

Non-local means label fusion:
The non-local means estimator is used to perform a weighted average of the labels based on the intensity distance between patches. The decision function $v(x_i)$ is given by:

$$v(x_i) = \frac{\sum_{s=1}^{N}\sum_{j\in V_i} w(x_i, x_{s,j}) y_{s,j}}{\sum_{s=1}^{N}\sum_{j\in V_i} w(x_i, x_{s,j})} \tag{5.2}$$

where $y_{s,j}$ is the manual annotation given to voxel $x_{s,j}$ at location j in subject s. $w(x_i, x_{s,j})$ is the weight assigned to $y_{s,j}$ by patch comparison. The weight is computed as:

$$w(x_i, x_{s,j}) = \begin{cases} exp^{\frac{-\|P(x_i)-P(x_{s,j})\|_2^2}{h^2}} & \text{if ss} > th \\ 0 & \text{else} \end{cases} \tag{5.3}$$

where $P(x_i)$ represents the cubic patch centered at x_i and $\| \, . \, \|_2^2$ is the normalized L2 norm (normalized by the number of elements) computed between each intensity element of patches $P(x_i)$ and $P(x_{s,j})$.

If the labels are considered to be binary, 0 corresponding to background and 1 to object, then:

$$L(x_i) = \begin{cases} 1 & v(x_i) > 0.5 \\ 0 & v(x_i) < 0.5 \end{cases} \tag{5.4}$$

h in Equation 5.3 is the decay parameter. When h is low only a few samples are taken into account, whereas a large value of h indicates that all samples have the same weight, and the estimation becomes a simple average. If patches very similar to the patch under study are estimated, h should be decreased to reduce the influence of other patches. When no similar patches are available, h should be increased to ensure that more patches are used in the label fusion. The estimation of $h(x_i)$ is done using:

$$h^2(x_i) = \lambda^2 \times \arg\min_{x_{s,j}} \parallel P(x_i) - P(x_{s,j}) \parallel_2^2 + \varepsilon \tag{5.5}$$

where ε is a small constant to ensure stability in case the patch under study is present in the training data. λ=0.5 as proposed in [7].

Figure 5.2 illustrates the patch-based segmentation.

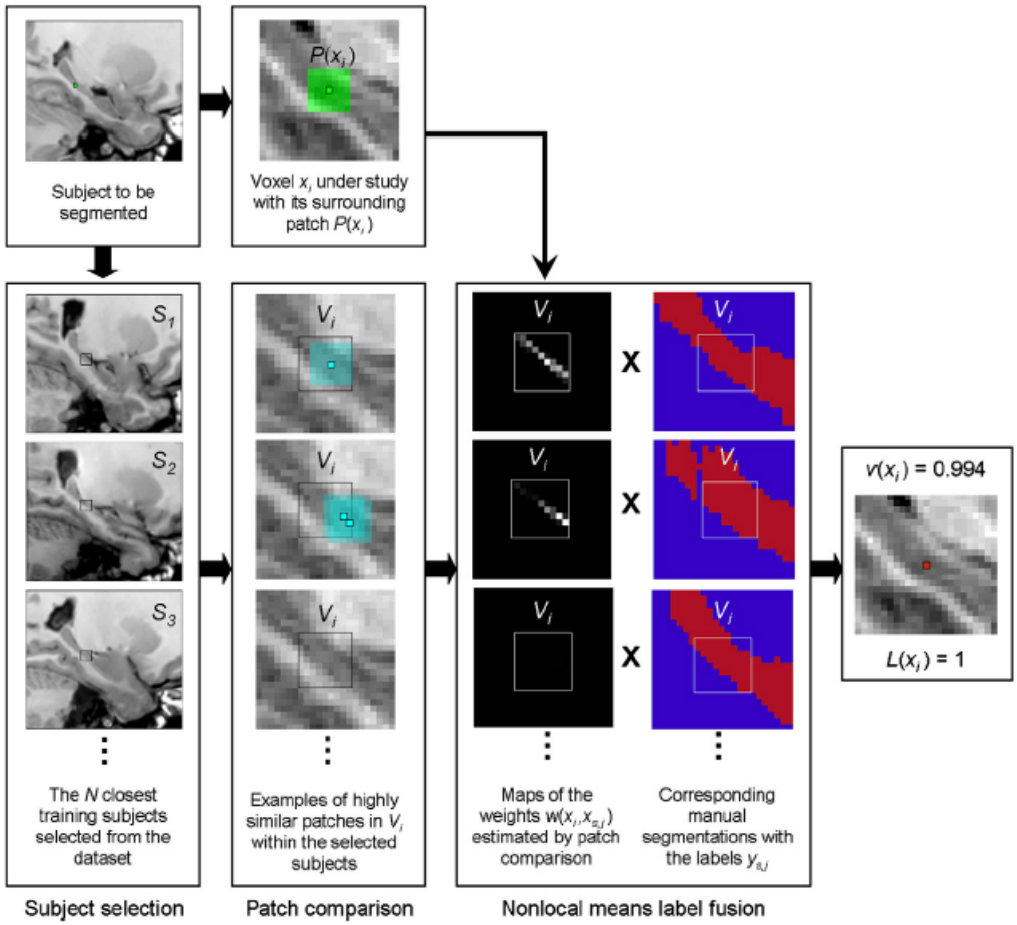


Figure 5.2: Overview of the Non-Local Patch-based segmentation. Segmentation of voxel x_i. The patch (green) is compared with all patches within the search volume of the N closest subjects. Highest weights are obtained by the most similar patches (blue) [8].

5.2 BrainFuseLab

As described in Chapter 2, many multi-atlas segmentation methods exist. In this thesis, BrainFuseLab (BFL) is chosen [28]. A test image is registered with each training subject using a diffeomorphic registration from ITK [16]. Using this transformation, the manual annotations are propagated to novel image coordinates approximately corresponding to the test subject's coordinates. Label fusion is reduced to a local weighted averaging, where training subjects that are locally more similar to the test subject in terms of intensity get more weight. The method is developed to use bias field corrected, intensity normalized, skull-stripped images with isotropic voxels preprocessed in FreeSurfer as input. The original code uses an atlas set with several subcortical structures. Therefore, the code has been changed slightly since only hippocampal labels are available in the HHP atlases.

Transfer of manual annotations:
The MRI is preprocessed in FreeSurfer as described in Chapter 4. To get the manual labels to FreeSurfer space, a rigid-body registration, T1, is computed, as in illustrated in Figure 4.6. This transformation is used to move the manual segmentations to FreeSurfer space using nearest neighbor interpolation.

Subject selection:
Initially, all atlases are registered to the test subject using an affine registration. Sums of squared differences (SSD) across an initialization mask (the skull-stripped brain mask) are calculated and the N closest subjects are selected. The affine parameters are saved and used later.

Non-rigid registration:
The non-rigid registration is an ITK-based implementation of a Demon's-based registration algorithm which can be found in [32]. In brief, this registration scheme a stationary velocity field (SVF) setting where paths of diffeomorphism are generated using one parameter subgroups through the Lie group exponential. The Lie group exponential is realized through a series of self compositions of a warp function. The warp Φ is parameterized with a smooth stationary field $v : R^3 \mapsto R^3$ via an Ordinary Differential Equation (ODE):

$$\frac{\partial \Phi(x,t)}{\partial t} = v(\Phi(x,t)) \tag{5.6}$$

where the warp is defined as $\Phi(x) = exp(v)(x)$ with v being the velocity field.

Since the unidirectional registration is asymmetric due to the integral over different volume forms, symmetry is ensured by transforming target volume form

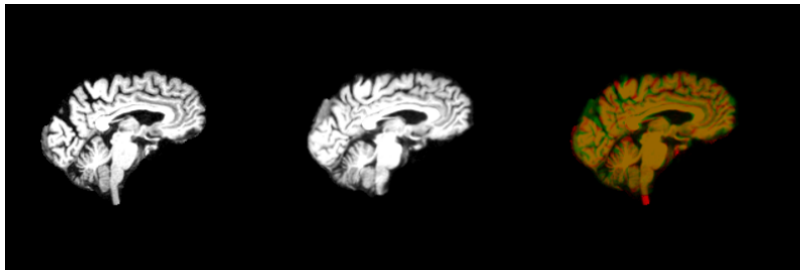
during the optimization using the Jacobian of the transformation. The following cost function is used to solve the variational problem to obtain an optimum velocity field:

$$\begin{aligned} \hat{v}^n = \arg\min_v \sum_{y\in\Omega} \Big[(I(y) - \tilde{I}_n(exp(v)(y))\Big]^2 \\ + \Big[(I(exp(-v)(y)) - \tilde{I}_n(y)\Big]^2 det(\nabla exp(-v)(y)) \\ +4\lambda\sigma^2 \sum_{j,k=1,2,3} \Big(\frac{\partial^2}{\partial x_j^2} v_k(x) \mid_{x=y}\Big)^2 \end{aligned} \tag{5.7}$$

where $\lambda > 0$ is the regularization parameter. x_j and v_k denotes the jth and kth dimension of the spatial position x, n is the n^{th} training image, v is the velocity, σ^2 is the stationary noise variance, I_n denotes the N training images and $\tilde{I}_n$ is the N training images where the spatial mapping from the test subject coordinates to the coordinates of the n^{th} training images, $\Phi_n : \Omega \mapsto R^3$, is unknown. Regularization is achieved by convolving the velocity field updates with a Gaussian: $K(x) \propto exp(-\alpha \sum_{n=1,2,3} x_i^2)$, where $\alpha = \gamma/8\lambda\sigma^2$ at every optimization step. α determines the smoothness of the final warp and $\gamma > 0$ controls the size of the Gauss-Newton step. Different values of γ are tried out in Chapter 6. Gauss-Newton scheme, section 4.2.3, is used to solve the ODE.

In Figure 5.3 a test subject and the corresponding closest training subject prior to and after warping the training image to the test image coordinates using the non-rigid registration can be seen.

Test Training Test+Training



Test Warped training Test+Warped

Figure 5.3: Subject 003_S_0931 and the training image before (row1) and after warping (row2). Red channel: Test subject. Green channel: Training image.

Local Weighted Voting Label Fusion:
The label fusion method is derived within a probabilistic framework. The goal is to estimate the label map L associated with the test image I, which can be achieved via a maximum-a-posterior (MAP) estimation.

$$\hat{L} = \arg\max_{L} p(L, I; \{L_n, I_n\}) \tag{5.8}$$

Where I_n denotes the N training images with corresponding label maps L_n, n = 1,...,N.

In the following $M : \Omega \mapsto \{1, \ldots, N\}$ denotes the latent random field that for each voxel in the test image I specifies the index of the training image I_n it was generated from. The image I and the label map L can be generated from

a mixture model, given a prior on M.

$$p(L, I; \{L_n, I_n\}) = \sum_M p(M) p(L, I \mid M; \{L_n, I_n\}) \tag{5.9}$$

It is assumed that each voxel is generated from a single training subject indexed with $M(x)$, i.e., $p(L(x) \mid M; \{L_n\}) = p_{M(x)}(L(x); L_{M(x)})$ and $p(I(x) \mid M; \{I_n\}) = p_{M(x)}(I(x); I_{M(x)})$. Inserting this into 5.9 gives:

$$p(L, I; \{L_n, I_n\}) = \sum_M p(M) \prod_{x \in \Omega} p_{M(x)}(L(x); L_{M(x)}) p_{M(x)}(I(x); I_{M(x)}) \tag{5.10}$$

The final cost function is achieved by substituting 5.10 into 5.8:

$$\hat{L} = \arg\max_L \sum_M p(M) \prod_{x \in \Omega} p_{M(x)}(L(x); L_{M(x)}) p_{M(x)}(I(x); I_{M(x)}) \tag{5.11}$$

Equation 5.11 has 3 individual terms, image likelihood ($p_{M(x)}(I(x); I_{M(x)})$), label prior ($p_{M(x)}(L(x); L_{M(x)})$) and membership prior $p(M)$. Variations in these terms gives different label fusion strategies.

For local weighted voting, $M(x)$ is independent and identically distributed according to an uniform distribution over all labels for all $x \in \Omega$, which means the membership prior becomes:

$$p(M) = \frac{1}{N^{|\Omega|}} \tag{5.12}$$

This reduces Equation 5.11, with $\mathcal{L}$ denoting the number of labels including background, to:

$$\hat{L}(x) = \arg \max_{l \in \{1, \ldots, \mathcal{L}\}} \sum_{n=1}^{N} p_n(L(x) = l; L_n) p_n(I(x); I_n) \tag{5.13}$$

The image likelihood serves as weights and is modeled as a Gaussian distribution with stationary variance σ^2:

$$p_n(I(x); I_n) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big[- \frac{1}{2\sigma^2} \Big(I(x) - \tilde{I}_n(\Phi_n(x)) \Big)^2 \Big] \tag{5.14}$$

The label prior term serves as votes and is given by:

$$p_n(L(x) = l; L_n) = \frac{1}{\sum_{l=1}^{\mathcal{L}} \left(\rho \tilde{D}_n^l(\Phi_n(x))\right)} exp\left(\rho \tilde{D}_n^l(\Phi_n(x))\right) \tag{5.15}$$

where $\tilde{I}_n$ is N training images where the spatial mapping from the test subject coordinates to the coordinates of the n^{th} training images, $\Phi_n : \Omega \mapsto R^3$, is unknown. $\tilde{D}_n^l$ denotes the distance transform of label l in training subject n. ρ is a slope constant.

Chapter 6

Parameter and method selection

To find the best method to segment *ADNI504*, the appropriate parameters used in Non-Local Patch-based segmentation (N-L Patch) and BrainFuseLab must be found. To find these parameters, *leave-one-out cross-validation* is done. In *leave-one-out cross-validation* (LOOCV) one single observation from a dataset is used as test data, and the remaining observations are used as training data. This is repeated until each observation in the dataset is used once as test data. Due to computational cost, LOOCV will initially be done on 15 atlases, *Atlas15*. When the appropriate parameters have been found, LOOCV will be done using *Atlas40*. ADNI504, Atlas15 and Atlas40 are explained in Chapter 3. Through this chapter cross-sectional FreeSurfer will be used as reference. Longitudinal FreeSurfer segmentations are not available for the atlases - only one time point scan is available for some of the atlases. Based on LOOCV, one method will be selected to segment ADNI504 and atrophy will be estimated and compared to cross-sectional and longitudinal FreeSurfer in Chapter 7. To compare the different methods and parameters, a volume overlap measure known as Dice score is used to evaluate the quality of the segmentations. Given an automatic segmentation $\hat{L}$ and the corresponding manual segmentation L, the Dice score of label l is given by [28]:

$$Dice(l;\hat{L},L) = 2\frac{|\{x \in \Omega | L(x) = l \& \hat{L}(x) = l\}|}{|\{x \in \Omega | L(x) = l\}| + |\{x \in \Omega | \hat{L}(x) = l\}|} \quad (6.1)$$

where $\Omega \subset R^3$ is a finite grid where the test subject is defined. The Dice scores varies between 0 and 1, where 0 indicates 0 % overlap with the manual segmentation and 1 indicates 100 % overlap with the manual segmentation - thus a perfect segmentation.

6.1 Atlas15 - Leave-one-out cross-validation

LOOCV with 15 atlases will be done on both N-L Patch and BrainFuseLab to find the appropriate parameters. These parameters will be applied in a LOOCV of Atlas40 in Section 6.2. The parameter annotation is the same as used in Chapter 5.

6.1.1 Non-Local Patch-based segmentation

N closest subjects:
Initially, a rigid-body registration is used to do both the intra- and inter-subject registration to one of the atlases in the atlas dataset, Figures 4.6 and 4.7. The labels and MRIs are transformed using the calculated transformation. The number of N closest atlases found under pre-selection can be varied from 1 to 14. The patch size and the search volume are set to 7×7×7 and 9×9×9, respectively, as suggested in [8]. Figure 6.1 illustrates mean Dice score after segmentation of 15 atlases as a function of a varying number of most similar atlases, N, from 1 to 14 after LOOCV. Dice scores are shown for both left and right hippocampus.

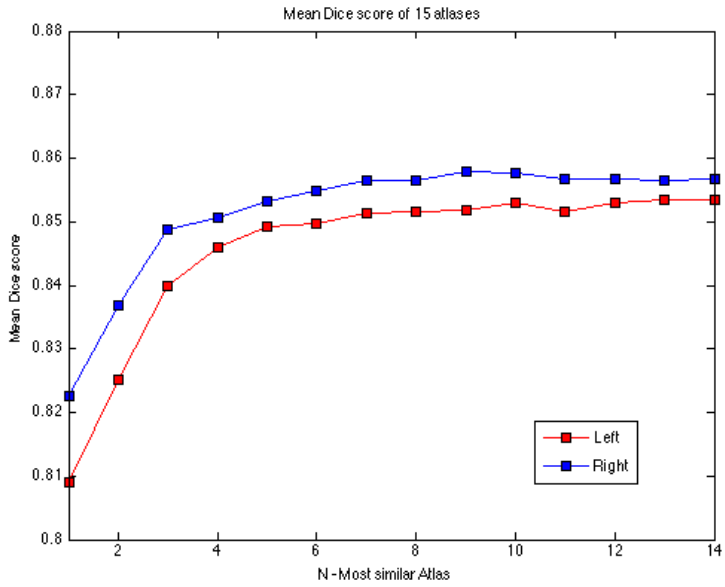


Figure 6.1: Mean Dice scores of 15 atlases as a function of varying number of most similar atlases, N, after leave-one-out cross-validation. Red: Left hippocampus. Blue: Right: hippocampus.

Figure 6.1 illustrates that after using approximately N=9 closest atlases in the segmentation, a steady state is reached. Therefore, N=9, will be used in N-L Patch.

Search volume and patch size:
The search volume (sv) can be viewed as the inter-subject variability of the structure. Since the hippocampus has a large variability, especially in pathological brains, one would expect that the search volume should be large compared to other structures with less variability. 3 different search volumes with side length 9, 11 or 13 are tested. For each search volume, 3 different patch sizes (ps) with side length, 3, 5 or 7 are tested. The Dice scores for left and right hippocampus with varying parameters can be seen in Table 6.1. N=9 closest subjects are used in the segmentation. The approximate segmentation computation duration in minutes per subject is shown in the table as well (Time).

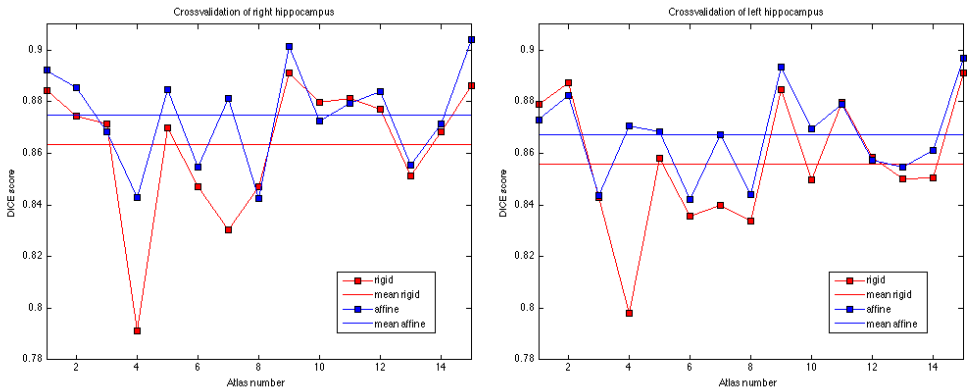
Search Volume		**9**	
Patch size	3	5	7
Dice score left $\pm\sigma$	0.829 ± 0.027	0.856 ± 0.026	0.853 ± 0.028
Dice score right $\pm\sigma$	0.836 ± 0.032	0.863 ± 0.027	0.858 ± 0.030
Time (min)	32	40	80
Search Volume		**11**	
Patch size	3	5	7
Dice score left $\pm\sigma$	0.825 ± 0.030	0.856 ± 0.025	0.854 ± 0.028
Dice score right $\pm\sigma$	0.831 ± 0.035	0.863 ± 0.027	0.858 ± 0.029
Time (min)	43	72	181
Search Volume		**13**	
Patch size	3	5	7
Dice score left $\pm\sigma$	0.816 ± 0.033	0.854 ± 0.025	0.852 ± 0.028
Dice score right $\pm\sigma$	0.821 ± 0.040	0.859 ± 0.028	0.856 ± 0.029
Time (min)	68	126	308

Table 6.1: Search volume and patch size impact on Dice score. A patch size of e.g. 3 corresponds to a $3\times3\times3$ volume. Dice scores for left and right hippocampus as well as the duration of segmenting both left and right hippocampus for one subject can be seen.

Based on both precision in terms of Dice score and time, a patch volume of side length 5 and a search volume of side length 9 will be used.

Affine vs. rigid registration:
The inter-subject registration can be done using a rigid-body registration or

an affine registration, Figure 4.7. To test the impact on the precision, both registrations are tried out in LOOCV, Figure 6.2. Mean Dice scores are denoted by the horizontal line in the figure and can be seen in Table 6.2 as well.



(a) Right (b) Left

Figure 6.2: Dice scores as a function of atlas number. Segmentation of the 15 atlases using inter-subject rigid registration (red) or affine registration (blue). LOOCV using N=9, sv=9, ps=5. Mean Dice scores are denoted by the horizontal line.

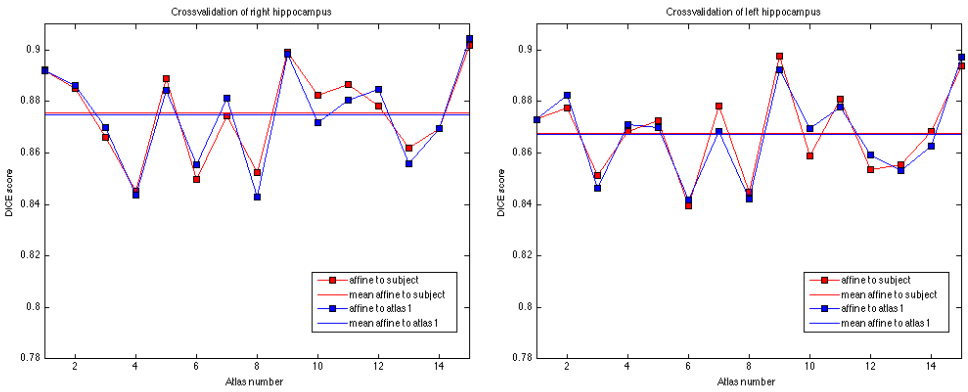
	Mean Dice score $\pm\sigma$	
	Right	Left
Affine	0.875±0.019	0.867±0.017
Rigid	0.863±0.026	0.856±0.025

Table 6.2: Mean Dice scores of LOOCV using 15 atlases where labels are aligned using affine or rigid registration.

Since an affine registration results in both a larger mean Dice score and a smaller standard deviation, Table 6.2, than using a rigid registration, affine registration will be used to do the inter-subject registration to transform labels and MRIs.

Align labels to test subject or standard atlas:
Two different ways in aligning the atlases and test subject to a segmentation space have been tested. Dice scores as a function of atlas number can be seen in Figure 6.3. In Figure 6.3 (a), both the test subject and the atlases are aligned to the first atlas in the atlas set, atlas1 (Affine to atlas1), blue. In Figure 6.3 (b), all the atlases are aligned to the test subject (Affine to subject), red. The

corresponding mean Dice scores are denoted by the horizontal line and is stated in Table 6.3.



(a) Right (b) Left

Figure 6.3: Dice scores as a function of atlas number. Red: Test subjects and atlases are all aligned to atlas1 using affine registration (affine to atlas1). Blue: Atlases are aligned to the test subject using affine registration (affine to subject).

	Mean Dice score $\pm\sigma$		
	Right	Left	Time (min)
Affine to atlas1	0.875±0.019	0.867±0.017	40
Affine to subject	0.876±0.018	0.868±0.017	60

Table 6.3: Mean Dice scores of LOOCV using 15 atlases where labels are aligned using affine transformation to either atlas1 (Affine to atlas1) or the test subject (Affine to subject). Furthermore, the computation time of segmenting both left and right hippocampus in a subject is stated (Time).

As Figure 6.3 and Table 6.3 illustrates, doing the inter-subject registration of all atlases to atlas1 once or doing the registration of all atlases to the test subject, gives the same results. Aligning all atlases to atlas1 only has to be done once. Segmentation of a new test subject then only requirers one registration, which takes the test subject's coordinates to atlas1's coordinates. This approach is considerably faster than aligning all atlases to the test subject each time. Therefore, aligning to atlas1 will be used onwards.

Threshold of non-local means estimator:
According to [8], the threshold of the non-local means estimator $v(x_i)$, Equation 5.4, decides if a voxel belongs to the object (L=1) or the background (L=0). This threshold is suggested to be 0.5 in [8]. To verify if this is the optimal value for this implementation, the threshold of $v(x_i)$ is varied from 0 to 1, which leads to a number of slightly different segmentations. The total Dice score (both left and right hippocampus) with the manual segmentations are calculated. The plot of Dice scores as a function of the threshold of $v(x_i)$ can be seen in Figure 6.4.

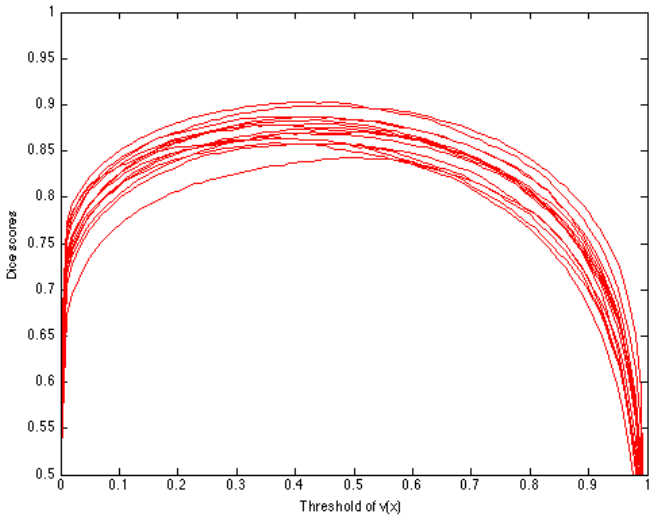


Figure 6.4: Dice scores as a function of varying threshold $v(x_i)$ of Atlas15. Each curve illustrates the behavior of one atlas.

The maximum Dice score and the corresponding threshold $v(x_i)_{max}$ of each atlas is found. The 15 $v(x_i)_{max}$ have a mean value of 0.42. This results in a mean Dice score of 0.875. A threshold of 0.5 as suggested in [8] results in a mean Dice score of 0.871. Since it is only the 3. decimal that is affected by changing the threshold from 0.42 to 0.5, a threshold of 0.5 will be used as suggested in [8].

Removal of small connected components:
Many segmentations have speckle patterns (black) as illustrated in row 2 in Figure 6.5. Therefore, a modification of the original method is made. Connected components are found using a 6-connectivity neighborhood and are removed if their total volume is less than 100 voxels. The Dice scores, before and after

removal can be seen in Table 6.4. The table indicates, that the removal does not affect Dice scores. However, a visually more satisfying result is achieved after removal, row 3 Figure 6.5, and therefore a removal of small connected components will be performed on top of N-L Patch.

	Mean Dice score$\pm\sigma$	
	Right	Left
Before removal	0.875±0.019	0.867±0.017
After removal	0.875±0.019	0.868±0.017

Table 6.4: Mean Dice scores of LOOCV using 15 atlases before and after removing connected components with volumes less than 100 voxels.

6.1.2 BrainFuseLab

BrainFuseLab has many parameters that can be varied in wide ranges, Section 5.2. According to [28] it is especially the following parameters which affect the result: γ that controls the step size in the gauss-newton optimization, and the standard deviation σ in the Gaussian Image Likelihood term used for label fusion, Equation 5.14. To make a fair comparison with N-L Patch, N=9 closest atlases will be used. The settings in the downloaded demo ($\gamma = 20, \sigma = 5$) and the default values ($\gamma = 150, \sigma = 10$), have been tested. LOOCV using the default and demo parameters can be seen in Figure 6.6. Notice, that the y-axis has been changed due to very low Dice-scores compared to Figure 6.3. The mean Dice scores $\pm\sigma$ can be seen in Table 6.5.

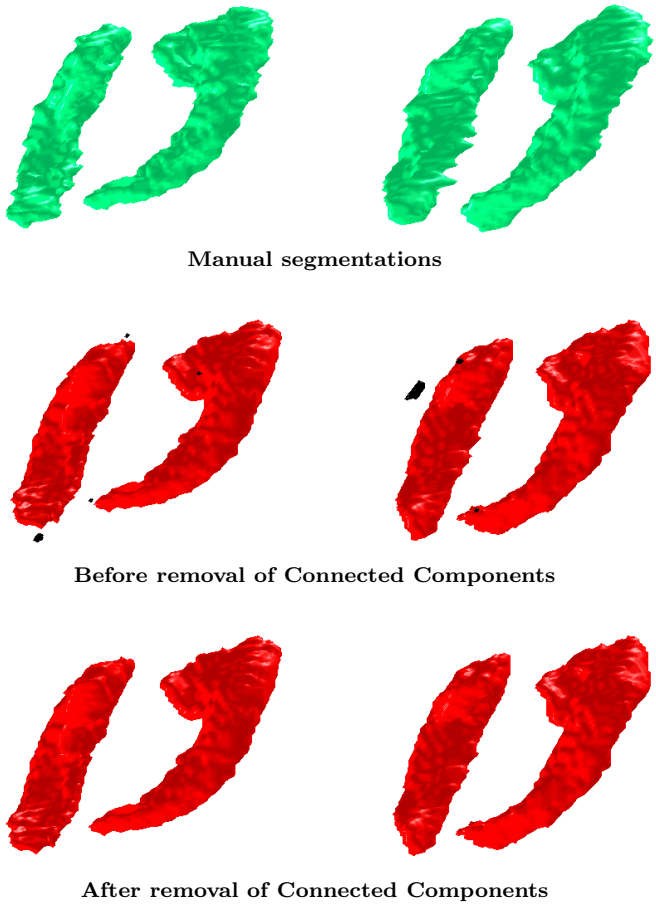
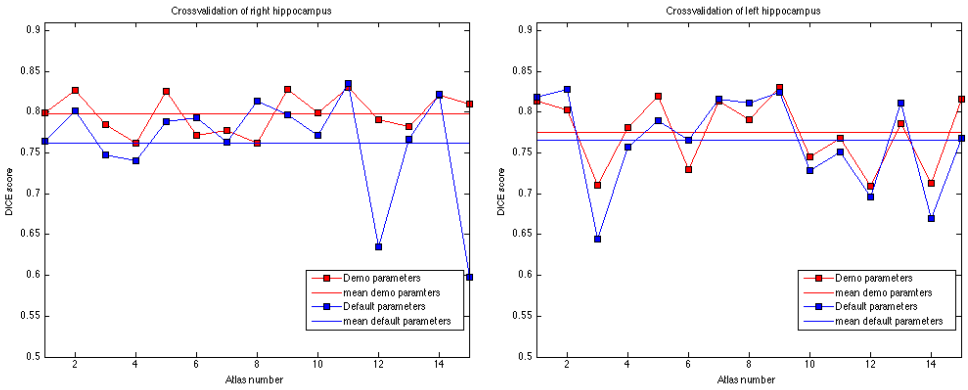


Figure 6.5: 3D illustrations of removal of connected components (black, row 2) from two subjects left and right column respectively. Top: Manual segmentations. Middle: Segmentations (red) with connected components < 100 voxels (black). Bottom: After removal of connected components.



(a) Right (b) Left

Figure 6.6: Dice scores as a function of atlas number. BrainFuse-Lab LOOCV with demo parameters (red) and default parameters (blue).

	Mean Dice score $\pm\sigma$	
	Right	Left
Demo parameters	0.798±0.025	0.775±0.043
Default parameters	0.763±0.066	0.765±0.058

Table 6.5: Mean Dice scores of LOOCV with BrainFuseLab. Demo parameters and default parameters are tested.

Based on these findings, the demo parameters will be used further on.

6.2 Atlas40 - Leave-one-out cross-validation

25 extra atlases are introduced and LOOCV is done. The optimal parameters found using Atlas15 are used to do N-L Patch segmentation as well as segmentation with BrainFuseLab (BFL). As suggested as future work in e.g. [8], the non-rigid registration from BrainFuseLab are combined with N-L Patch. The transformed labels for N closest atlases using BrainFuseLab are collapsed into a mask that serves as initialization mask for N-L Patch. The segmentation of both left and right hippocampus is done using the N closest atlases found in BrainFuseLab.
The Dice scores with the manual HHP labels, false positive error (FP) and false negative error (FN) with N-L patch, BFL, combination of non-rigid registration from BFL and N-L patch (Non rigid + N-L Patch) and cross-sectional FreeSurfer (FS Cross) can be seen Table 6.6. Furthermore, a box plot of the Dice scores of the different methods can be seen in Figure 6.7. When calculating Dice scores, there has not been distinguished between right and left hippocampus as done in most of section 6.1.
Notice that the number of atlases and number of closest atlases are fundamentally different for cross-sectional FreeSurfer, Table 6.6. Cross-sectional FreeSurfer uses a probabilistic atlas build from 39 atlases as explained in Section 3.2.2, which means averaged information from all these atlases are used in the segmentation. The FreeSurfer Dice scores are obtained by calculating the overlap of segmentations with the HHP labels, even though FreeSurfer uses another atlas to do segmentation with a different definition of hippocampus. By doing this, the FreeSurfer segmentation consensus with the new label standard is illustrated.

Method	Atlas	N	Mean Dice score $\pm\sigma$	Mean FP $\pm\sigma$	Mean FN $\pm\sigma$
N-L Patch	40	9	0.868 ± 0.019	0.125 ± 0.025	0.138 ± 0.028
BFL	40	9	0.827 ± 0.027	0.139 ± 0.040	0.203 ± 0.034
Non rigid + N-L Patch	40	9	0.857 ± 0.016	0.106 ± 0.021	0.177 ± 0.023
FS Cross	39	-	0.781 ± 0.031	0.158 ± 0.030	0.270 ± 0.052

Table 6.6: Atlas40 LOOCV. Atlas: Number of atlases available during preselection based on SSD. N closest: Number of closest atlases based on SSD used in segmentation. FP: False Positive error. FN: False negative error.

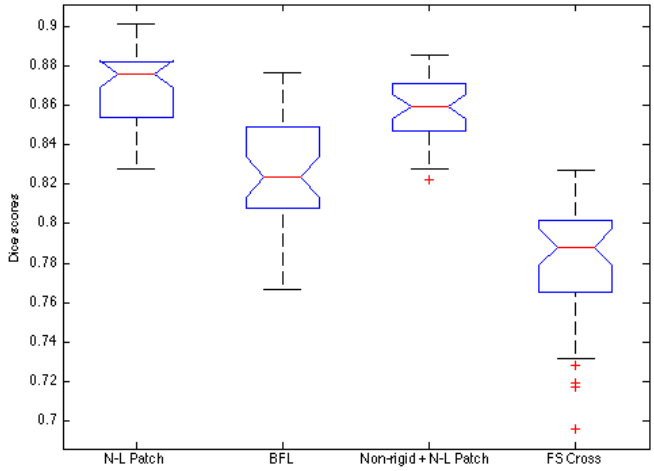


Figure 6.7: Box plot of the Dice scores with different methods. Boxes represents the lower quartile, the median (red line) and the upper quartile. Whiskers indicate the extreme values within 1 times the interquartile range. Outliers (red +) are the data values beyond the end of the whiskers.

As seen in Table 6.6, the best mean Dice score is achieved using N-L patch. However, the mean Dice score is approximately the same as the one achieved using 15 atlases (mean Dice score = 0.871). On the other side, the mean Dice score has increased for BFL. This indicates, that N-L patch due to the fact that it can use more than one similar patch from each atlas, do not need as many available atlases in order to perform well as methods that only uses information once for each atlas to segment a voxel. It should further be noticed, that combining the non-rigid registration from BrainFuseLab and N-L patch increases the Dice score compared to BFL from a mean of 0.827 $\pm$ 0.027 to 0.857 $\pm$ 0.016. This illustrates N-L Patch's label fusion capabilities compared to the local weighted voting label fusion used in BrainFuseLab.

The approximate computation times can be seen in Figure 6.7. The codes are all a single core CPU implementation on a 2.5 GHz Xeon.

Method	Computational time (hours)
N-L Patch	~ 0.7
BFL	~ 5
Non rigid + N-L Patch	~ 5.5
FS Cross	~ 11

Table 6.7: Computational times for each subject with different methods. It should be noticed that FS Cross segments 37 subcortical structures incl. the hippocampi, whereas the other methods only segments two: left and right hippocampus.

Statistics:
Paired t-tests between methods based on Dice scores from LOOCV using 40 atlases have been made independent of CN, MCI and AD, Table 6.8. The p-values are < 0.001 for N-L patch vs. all other methods, which according to the null hypothesis means that equal means can be rejected.

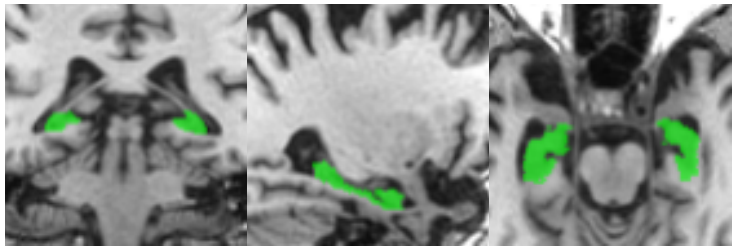
Methods	**t-value**	**p-value**
N-L Patch vs. FS Cross	27.6346	<0.001
N-L Patch vs. BFL	13.5424	<0.001
N-L Patch vs. Non rigid + N-L Patch	7.6142	<0.001
BFL vs. FS Cross	12.9467	<0.001
BFL vs. Non rigid + N-L Patch	-9.892	<0.001
FS Cross vs. Non rigid + N-L Patch	-21.1403	<0.001

Table 6.8: Results of paired t-tests between Dice scores of different methods.

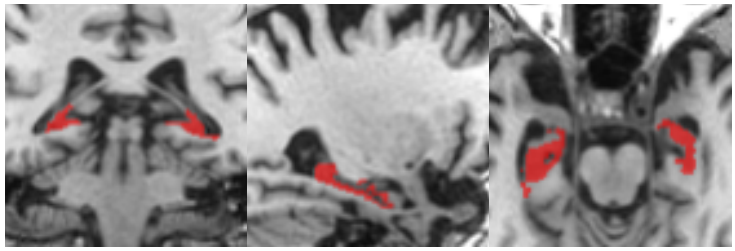
Illustrations:
Based on the FreeSurfer segmentations, the worst, median and best atlas subject is selected. The segmentations of these three subjects segmented using different methods are illustrated in a coronal, sagittal and transversal view, Figures 6.8, 6.9 and 6.10. Notice, that in order to illustrate the N-L Patch segmentations in the same segmentation space as the other methods, they must be taken back to the subjects FreeSurfer space, by applying the inverse transformation, $T2^{-1}$, from Figure 4.7. This involves a nearest neighbor interpolation. Due to this interpolation, the mean Dice score changes from 0.868 ± 0.019 to 0.855 ± 0.021. However, as illustrated in Table 4.1 this only changes volume size a little. Hippocampal volume size will be used to calculated atrophy. Furthermore, the 3D segmentations of the worst, median and best subject can be seen in Figure 6.11.

From Figures 6.8, 6.9, 6.10 and 6.11 it can be seen that both N-L Patch and the non-rigid registration combined with N-L Patch, visually look most like the manual segmentations (green). This is also reflected in the Dice scores achieved with these methods. The FreeSurfer segmentations are rough, whereas the BranFuseLab segmentations have some speckles after segmentation both at the hippocampal borders as well as within the background labels.

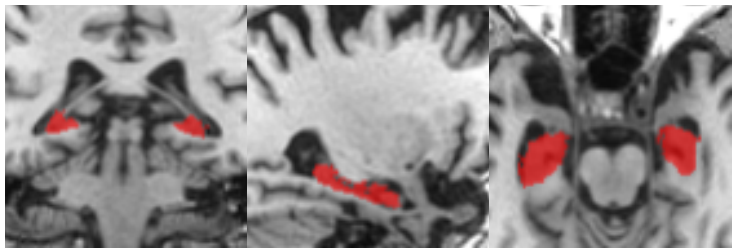
Figure 6.8: Worst subject 002_S_0938 (AD)



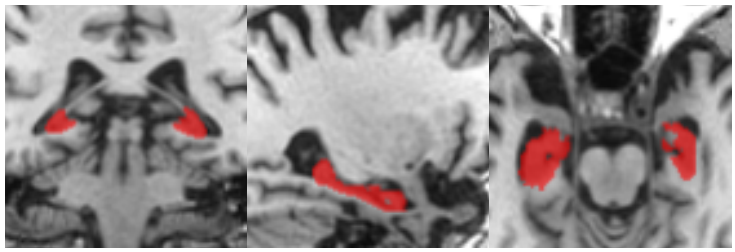
Manual segmentation



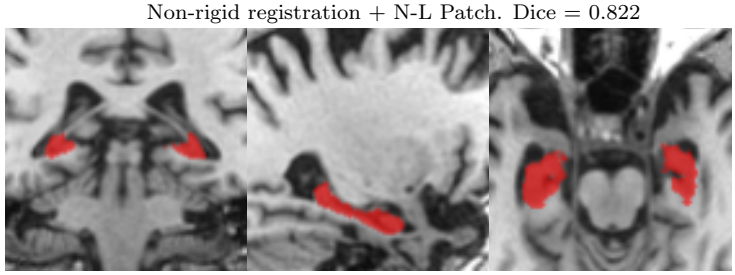
FreeSurfer. Dice = 0.696



BrainFuseLab. Dice = 0.766

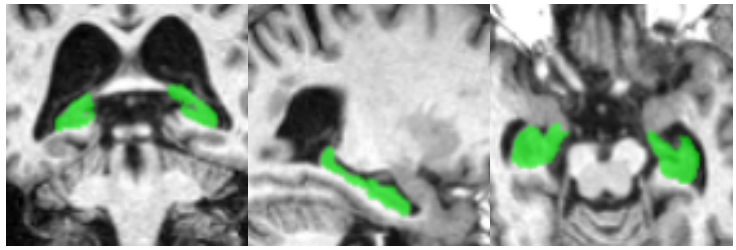


Non-rigid registration + N-L Patch. Dice = 0.822

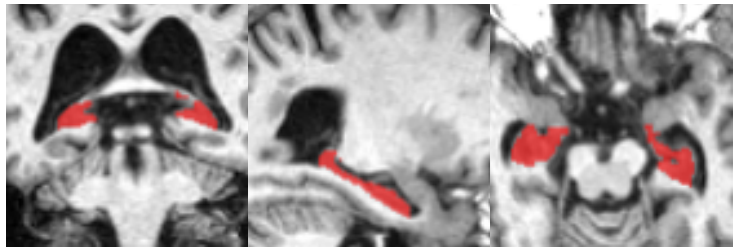


N-L Patch. Dice = 0.821

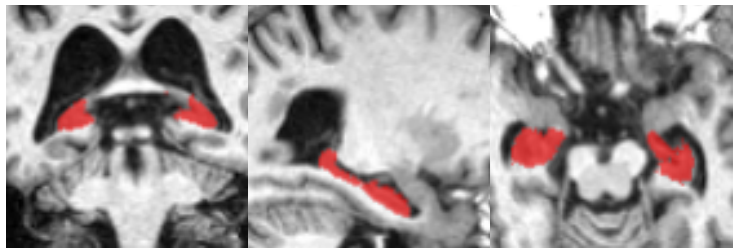
Figure 6.9: Median subject 013_S_1276 (CN)



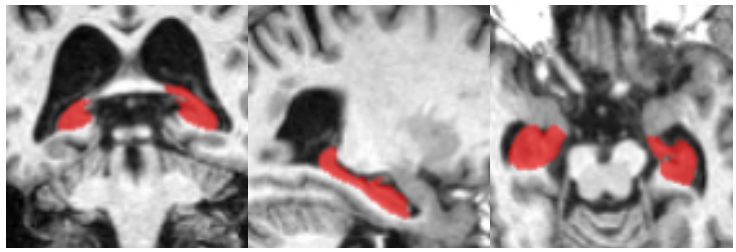
Manual segmentation



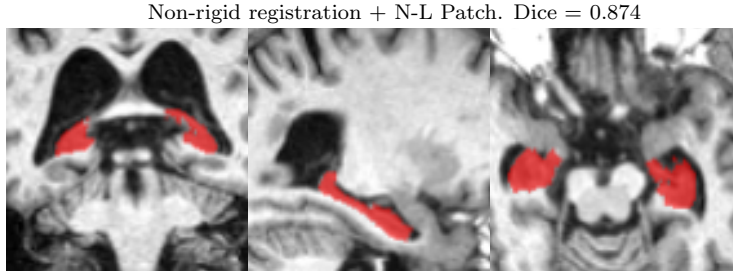
FreeSurfer. Dice = 0.788



BrainFuseLab. Dice = 0.847

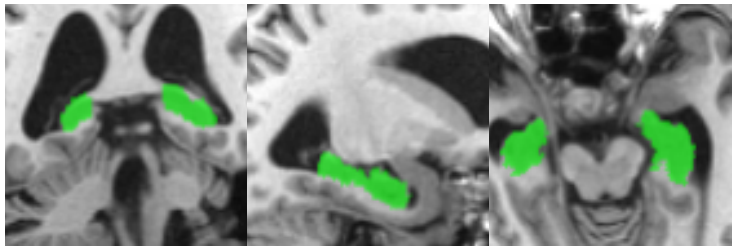


Non-rigid registration + N-L Patch. Dice = 0.874

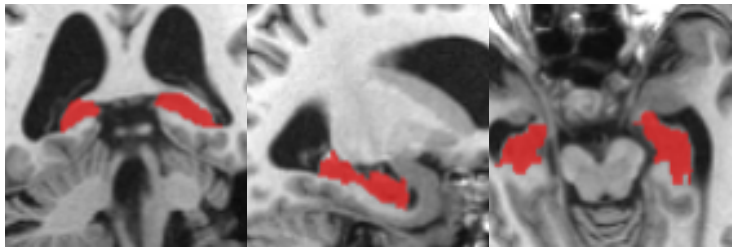


N-L Patch. Dice = 0.871

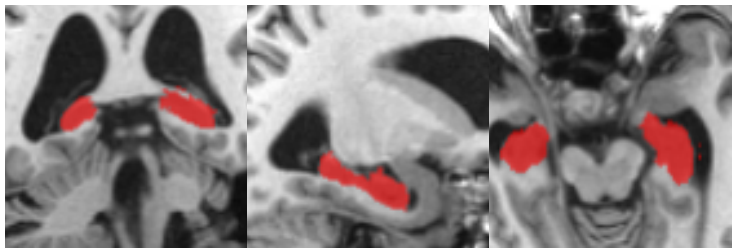
Figure 6.10: Best subject 011_S_0002 (CN)



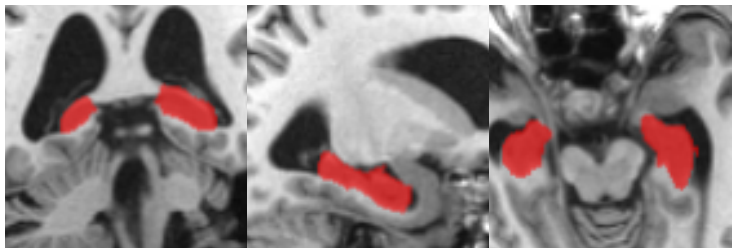
Manual segmentation



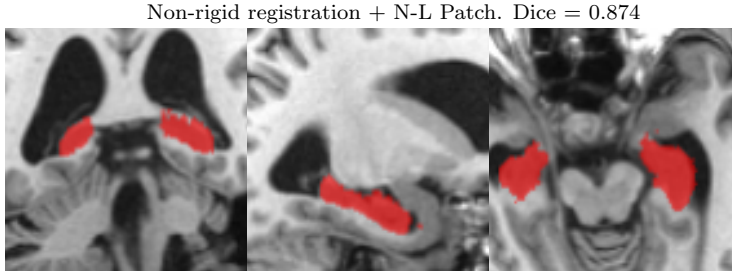
FreeSurfer. Dice = 0.827



BrainFuseLab. Dice = 0.867



Non-rigid registration + N-L Patch. Dice = 0.874



N-L Patch. Dice = 0.875

Figure 6.11: 3D illustrations of left: Worst subject. Middle: Median subject. Right: Best subject. Selected based on FreeSurfer LOOCV results.

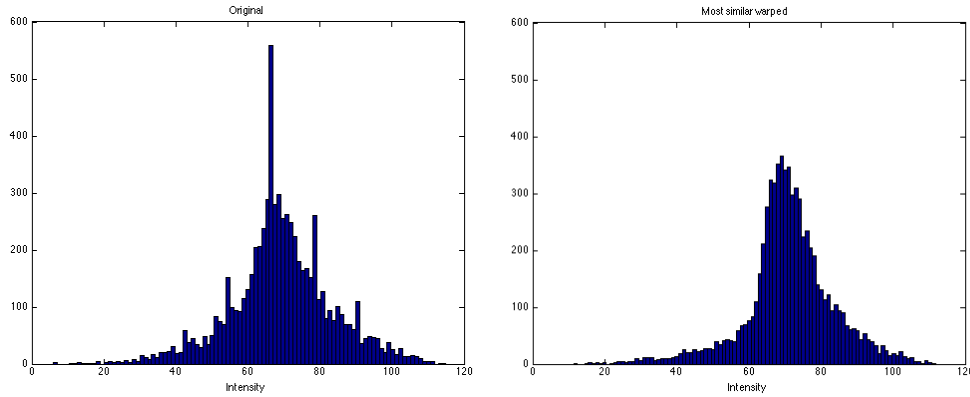
6.3 Evaluation

N-L Patch results in Dice scores significantly higher than all other methods using paired t-tests between methods independent of diagnostic groups, Table 6.8. Furthermore, N-L Patch is the fastest method. Therefore, N-L Patch will be implemented in Chapter 7.

Even though Dice scores are not improved by using 40 atlases compared to 15, 40 atlases will be used since this increases the library used for atlas preselection. Furthermore, the computation time is not increased more than a couple of minutes using 40 atlases compared to 15 atlases, since the time is depending most on the number of N closest atlases used in segmentation.

In [8] the best mean Dice score is reported to be 0.884. However, this score is from LOOCV of 80 healthy young subjects (mean age: 25.09 ± 4.9 years) and N=30 closest subjects. In this work a mean Dice score of 0.868 is achieved from LOOCV of 40 subjects with pathological processes (mean age: 74.10 ± 7.67 years) and N=9 closest subjects. The two numbers cannot directly be compared since it is more difficult to segment brains of elderly people with pathological processes then healthy young subjects due to large variability in the subjects.

Combining the non-rigid registration from BrainFuseLab with N-L Patch results in a mean Dice score of 0.857 ± 0.016 compared to BrainFuseLab with a mean of 0.827 ± 0.027. This indicates that the label fusion technique in N-L Patch is better than the technique used in BrainFuseLab. Whether this is due to a more comprehensive parameter optimization in N-L Patch or the fact that N-L Patch also uses information from shifted voxels is hard to tell. Figure 6.12 illustrates the histogram of one subject inside the manual hippocampal mask (left) and the corresponding histogram of the most similar atlas image (right) after transformation of the atlas using the non-rigid registration in BrainFuseLab. Large differences can not be observed between the histograms. This indicates that the mediocre results obtained with BrainFuseLab cannot be explained by the non-rigid registration.



(a) Test (b) Warped

Figure 6.12: Histogram inside hippocampusmask. Left: Test subject. Right: Most similar warped atlas after non-rigid registration.

Chapter 7

Final results

In Chapter 6, Non-Local Patch-based segmentation (N-L Patch) was found to be the most optimal method based on both precision (Dice scores) and computation time. Therefore, N-L Patch will be used to segment *ADNI504* at 3 time points, month 0 (baseline), month 12 (m12) and month 24 (m24). Atrophy will be estimated between baseline-month 12 (bam12) and baseline-month 24 (bam24). Since manual segmentations are only available for very few MRIs in ADNI504, no segmentation ground truth is available for the entire dataset, and Dice scores cannot be calculated. The evaluation of the method will be based on its diagnostic group separation capabilities compared to current methods used at *Biomediq A/S*, cross-sectional FreeSurfer and longitudinal FreeSurfer.

In this chapter, adjustments in preprocessing, inter-subject registration and skull-stripping, is done for some subjects to achieve the final segmentations of ADNI504. Using the final segmentations, atrophy will be estimated and a statistical analysis will be made to evaluate the diagnostic group separation capabilities of N-L Patch compared to cross-sectional FreeSurfer and longitudinal FreeSurfer. Finally, the results are discussed.

In this thesis, longitudinal FreeSurfer segmentations are calculated exploiting two time points simultaneously (baseline and m12 or baseline and m24) to segment one time point. This means, that the baseline volume used to calculate bam12 is not the exact same as the baseline volume used to calculate bam24.

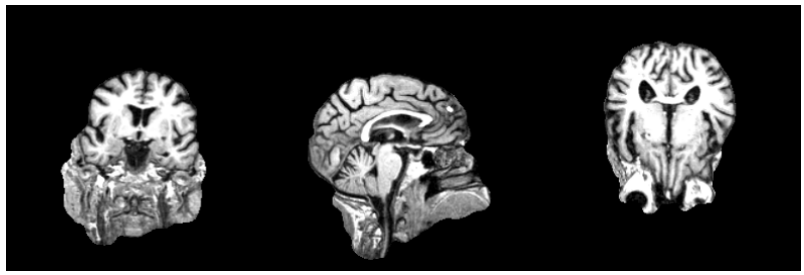
7.1 Method

N-L Patch segmentation of ADNI504 is done with search volume = 9, patch size = 5, number of atlases = 40, N closest atlases = 9 using an affine inter-subject registration to one atlas as suggested in Chapter 6. 21 atlases in Atlas40 are also part of ADNI504. When these 21 subjects are segmented their atlas is not used in the segmentation. Since a segmentation of one subject takes about 40 minutes, the total computation time used to segment all subjects was approximately 1000 hours.
After the initial segmentation, the hippocampal volume is estimated. If the total subject hippocampal volume is below 3500 voxels, a rigid-body inter-subject registration is done instead of the affine and the segmentation is done again. A segmentation volume of 3500 voxels is considered to correspond to an extremely insufficient segmentation, since the total hippocampal volume typically is 6500-8500 mm^3, as referenced in Table 3.2 and 3.3.
Since rigid registrations does not involve shearing and scaling, the registrations are not as sensitive to e.g. insufficient skull-strippings or enlarged ventricles as affine registrations. However, applying a rigid registration does not give good results in all cases. If segmentation is still very poor (total hippocampal volume below 3500 voxels and visual inspection) the preprocessed MRIs are inspected. In all cases, the poor results were due to an insufficient skull-stripping. The skull-stripping is changed by adjusting the watershed parameter, Section 4.1.2, until the MRI is correctly separated into brain tissue and non-brain tissue. The initial segmentation using affine registration is done again. Table 7.1 indicates the number of times the registration is changed or skull-stripping is redone to achieve the final results for ADNI504 at baseline, m12 and m24, respectively. After fixing the skull-stripping good results were achieved in all cases.

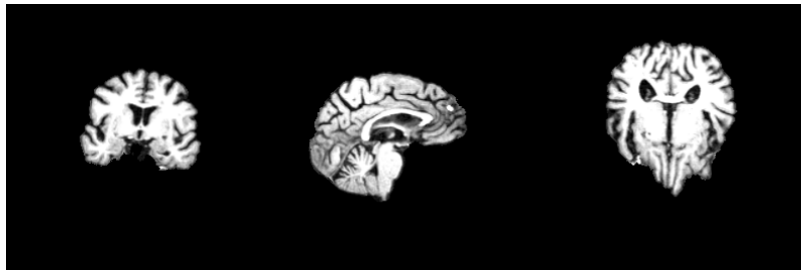
	Rigid	**Skull-strip**
Baseline	6	0
m12	12	3
m24	12	4

Table 7.1: Number of times registration fails and either a rigid registration is done or the skull-stripping is fixed. **Rigid:** If the total hippocampus volume was below 3500 voxels after the initial segmentation, then the inter-subject registration was changed to a rigid registration and the segmentation was done again. **Skull-strip:** If the result still was not satisfying after a rigid registration, the skull-stripping was changed and the segmentation was done again.

Figure 7.1 illustrates the MRI of subject 109_S_1114 m12 before and after fixing an insufficientt skull-stripping and Figure 7.2 illustrates the corresponding segmentations.



Initial skull-stripping



Final skull-stripping

Figure 7.1: Coronal, sagittal and transversal view of subject 109_S_1114 m12 before (row1) and after (row2) an insufficient skull-stripping has been fixed.

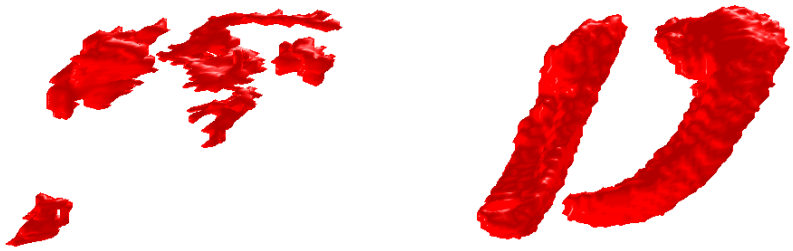


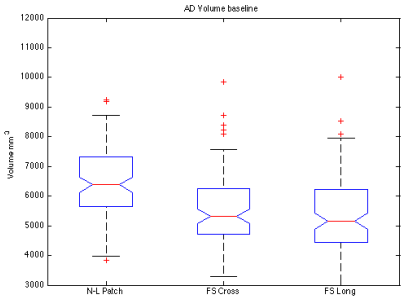
Figure 7.2: Hippocampal segmentation of subject 109_S_1114 m12 before (left) and after (right) an insufficient skull-stripping has been fixed, Figure 7.1.

To achieve the final segmentations, affine inter-subject registration was used for most subjects, whereas rigid inter-subject registration was used for a minority of subjects when the affine registration failed. Since affine registration involves scaling, the results achieved using an inter-subject affine registration cannot volumewise be compared to the results achieved using an inter-subject rigid registration. Thus, all segmentations are transformed back to the subject's FreeSurfer space, by applying the inverse transformation, $T2^{-1}$, from Figure 4.7, before the the total segmentation volume is estimated.

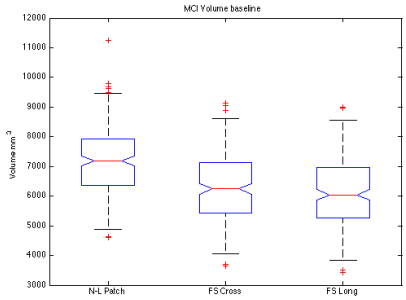
7.2 Segmentation results

The sum of voxels (1 voxel = $1mm^3$) for all 504 subjects at baseline, m12 and m24 for N-L patch, cross-sectional FreeSurfer and longitudinal FreeSurfer can be found on the CD in Appendix C. The corresponding volume box plots with the three methods at baseline can be seen in Figure 7.3 for AD, MCI and CN. The longitudinal volumes are calculated exploring baseline and m12 simultaneously. The mean $\pm$ σ can be seen in Table 7.2. A corresponding table containing m12 and m24 mean $\pm\sigma$ for diagnostic groups with N-L Patch, cross-sectional FreeSurfer and Longitudinal FreeSurfer can be seen in Appendix B. Furthermore, scatter plots of baseline hippocampal volume with N-L Patch against cross-sectional FreeSurfer, Figure 7.4, and N-L Patch against longitudinal FreeSurfer, Figure 7.5, can be seen. The scatter plots for N-L patch vs. cross-sectional FreeSurfer and longitudinal FreeSurfer for m12 and m24 can be

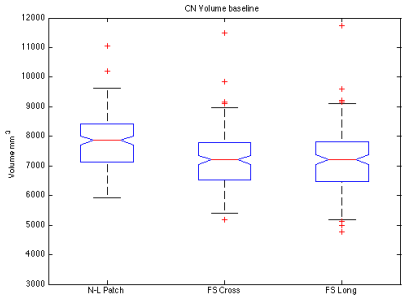
seen in Appendix B. The Figures and Tables illustrate that the volume with N-L patch is generally larger than the volume obtained with the FreeSurfer methods. This is most likely due to the fact that N-L Patch uses another atlas than FreeSurfer.



(a) AD



(b) MCI



(c) CN

Figure 7.3: Boxplot of baseline volume for AD, MCI and CN with three methods: N-L Patch, FS Cross and FS Long. Boxes represents the lower quartile, the median (red line) and the upper quartile. Whiskers indicate the extreme values within 1 times the interquartile range. Outliers (red +) are the data values beyond the end of the whiskers.

	Volume $(mm^3) \pm \sigma$		
	CN(n=169)	MCI(n=234)	AD(n=101)
N-L Patch	7860 ± 880	7166 ± 1083	6520 ± 1190
FS Cross	7210 ± 968	6313 ± 1263	5569 ± 1216
FS Long	7168 ± 1026	6117 ± 1083	5356 ± 1277

Table 7.2: ADNI504 baseline: Hippocampal volume for diagnostic groups with different methods.

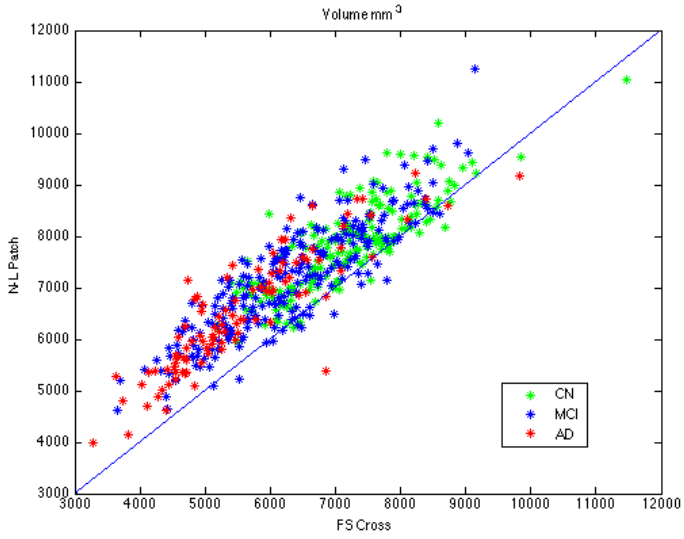


Figure 7.4: ADNI504 baseline: Hippocampal volume. N-L Patch vs. FS Cross.

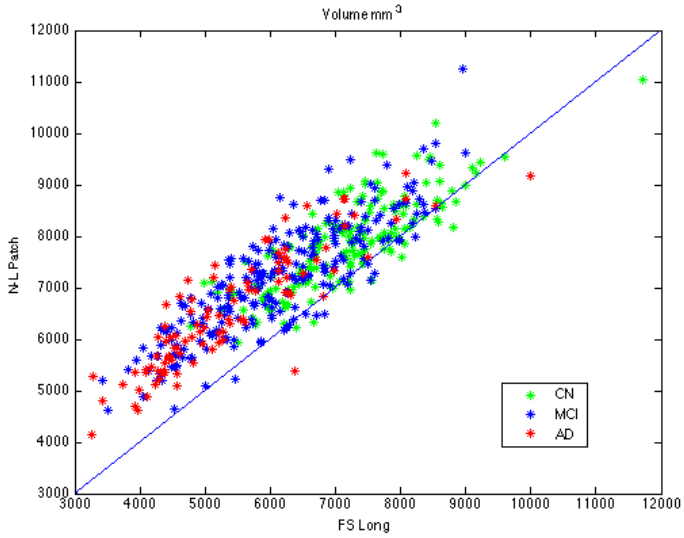


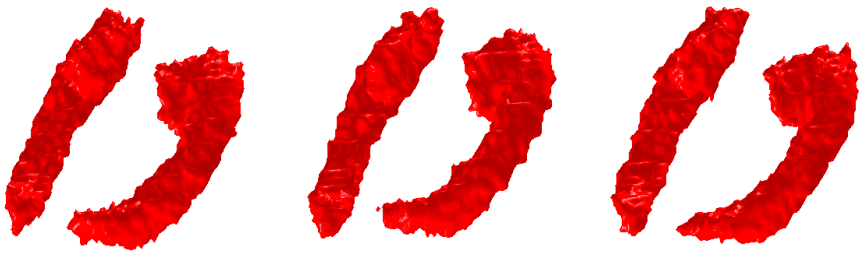
Figure 7.5: ADNI504 baseline: Hippocampal volume. N-L Patch vs. FS Long.

7.3 Statistical analysis

Based on the volume results, atrophy rate is estimated as percentage volume change from baseline to m12 (bam12) and baseline to m24 (bam24).

$$\text{atrophy}(\%) = \left(\frac{\text{Follow-up volume}}{\text{Baseline volume}} - 1\right) * 100 \tag{7.1}$$

The corresponding atrophy histograms with 50 bins for N-L Patch, cross-sectional FreeSurfer and longitudinal FreeSurfer can be seen in Appendix B. The segmentations of a MCI subject, 130_S_0783, with N-L Patch at baseline, m12 and m24 can be seen in Figure 7.6. Bam12 atrophy = -2.15 % and bam24 atrophy = -7.53 %. The subject is diagnosed with MCI at all 3 time points.



(a) Baseline (b) m12 (c) m24

Figure 7.6: Segmentation of subject 130_S_0784 (MCI) with N-L Patch at baseline (volume = 7234 mm^3), m12 (volume = 7044 mm^3) and m24 (volume = 6689 mm^3). Bam12 atrophy = -2.15 % and bam24 atrophy = -7.53 %.

Figure 7.6 illustrates the difficulty in visually seeing differences in volume between time points, even for this MCI subject with pathological changes in the brain.

Based on the atrophy scores, a statistical analysis will be made to evaluate the diagnostic group separation capabilities of N-L Patch compared to cross-sectional and longitudinal FreeSurfer. The statistical analysis of each method between diagnostic groups are done by performing two sample t-tests, AUCs and Cohen's Ds. Furthermore, an analysis between methods are done by comparing Cohen's Ds and AUCs. Bootstrapping is used to compare Cohen's Ds between methods, whereas DeLong test is used to compare AUCs between methods.

Initially, Bartlett's test of variance inhomogeneity is made and can be found in Appendix B. The low p-values reveal variance inhomogeneity between the clinical diagnostic groups in N-L Patch and longitudinal FreeSurfer for both bam12 and bam24 and cross-sectional FreeSurfer bam24. Variance homogeneity cannot be rejected for cross-sectional FreeSurfer bam12. This will be considered, when performing two sample t-tests between groups. The statistics calculated based on atrophy bam12 and bam24 can be seen in Tables 7.3 and 7.4. A brief analysis of the results will be given in the following sections.

	AD	**MCI**	**CN**
	mean $\pm\sigma$	mean $\pm\sigma$	mean $\pm\sigma$
N-L Patch	-4.23 $\pm$3.06	-2.39 $\pm$3.28	-0.86 $\pm$2.46
FS Cross	-4.29 $\pm$5.32	-3.69 $\pm$5.48	-1.39 $\pm$5.41
FS Long	-4.83 $\pm$3.74	-3.25 $\pm$3.53	-1.63 $\pm$2.54
	AD vs. CN	**AD vs. MCI**	**MCI vs. CN**
	t-test (p-value)	t-test (p-value)	t-test (p-value)
N-L Patch	-9.36 (<0.001)	-4.93 (<0.001)	-5.33 (<0.001)
FS Cross	-4.67 (<0.001)	-0.92 (0.357)	-4.16 (<0.001)
FS Long	-8.35 (<0.001)	-3.69 (<0.001)	-5.08 (<0.001)
	AD vs. CN	**AD vs. MCI**	**MCI vs. CN**
	AUC (p-value)	AUC (p-value)	AUC (p-value)
N-L Patch	0.80** (<0.001)	0.66*** (<0.001)	0.65 (<0.001)
FS Cross	0.69 (<0.001)	0.53 (0.404)	0.67 (<0.001)
FS Long	0.76 (<0.001)	0.62 (<0.001)	0.64 (<0.001)
	AD vs. CN	**AD vs. MCI**	**MCI vs. CN**
	Cohens'D	Cohens'D	Cohens'D
N-L Patch	1.21***	0.58**	0.53
FS Cross	0.54	0.11	0.42
FS Long	1.00	0.44	0.53

Table 7.3: Statistics based on atrophy (%) of ADNI504 between baseline and month 12. The first value for each t-test is the t-value. N-L Patch: DeLong test is done to compare AUCs between methods and bootstrapping is done to compare Cohens' Ds between methods. ** indicates significance of N-L Patch over FS Cross with p-value: 0.001$\leq$p-value<0.01; and *** indicates significance of N-L Patch over FS Cross with p-value<0.001.

	AD	**MCI**	**CN**
	mean $\pm\sigma$	mean $\pm\sigma$	mean $\pm\sigma$
N-L Patch	-7.55 $\pm$4.97	-4.54 $\pm$4.68	-1.86 $\pm$2.61
FS Cross	-9.40 $\pm$6.72	-7.17 $\pm$6.34	-3.15 $\pm$5.24
FS Long	-10.35 $\pm$5.11	-7.24 $\pm$5.85	-3.16 $\pm$3.16
	AD vs. CN	**AD vs. MCI**	**MCI vs. CN**
	t-test (p-value)	t-test (p-value)	t-test (p-value)
N-L Patch	-10.60(<0.001)	-5.16 (<0.001)	-7.28 (<0.001)
FS Cross	-7.96 (<0.001)	-2.81 (0.005)	-6.94 (<0.001)
FS Long	-12.71 (<0.001)	-4.87 (<0.001)	-8.98 (<0.001)
	AD vs. CN	**AD vs. MCI**	**MCI vs. CN**
	AUC (p-value)	AUC (p-value)	AUC (p-value)
N-L Patch	0.86 (<0.001)	0.69* (<0.001)	0.71 (<0.001)
FS Cross	0.82 (<0.001)	0.62 (<0.001)	0.73 (<0.001)
FS Long	0.89 (<0.001)	0.67 (<0.001)	0.73 (<0.001)
	AD vs. CN	**AD vs. MCI**	**MCI vs. CN**
	Cohens'D	Cohens'D	Cohens'D
N-L Patch	1.43*	0.62*	0.70
FS Cross	1.04	0.34	0.69
FS Long	1.69	0.57	0.87

Table 7.4: Statistics based on atrophy (%) of ADNI504 between baseline and month 24. The first value for each t-test is the t-value. DeLong test is done to compare AUCs between methods and bootstrapping is done to compare Cohens' Ds between methods. * indicates significance of N-L Patch over FS cross with p-value, $0.01 \leq$p-value<0.05.

Two sample t-tests:
For each method, two sample t-tests between two diagnostic groups at a time is performed to test the hypothesis:

$$\begin{aligned} H_0 &: \mu_1 - \mu_2 = 0 \\ H_1 &: \mu_1 - \mu_2 \neq 0 \end{aligned} \tag{7.2}$$

Based on the results of Bartlett's test, Appendix B, only variance homogeneity between groups will be assumed for cross-sectional FreeSurfer bam12. The t-values and p-values can be seen in Tables 7.3 and 7.4. Cross-sectional FreeSurfer AD vs. MCI bam12 is the only test, that is not significant. Thus there is a significant difference in atrophy rates for all other test scenarios.

AUC:
Area Under a ROC Curve (AUC) is defined as the area under a Receiver Operating Characteristics (ROC) graph. ROC graphs are commonly used in medical decision making to visualize a classifiers performance. The curve is achieved by changing the threshold of the measure which is evaluated, and for each threshold, calculate the false positive rate (x-axis) and the true positive rate (y-axis). (0,1) corresponds to a perfect classification. y=x represent randomly guessing a class. AUC will always be between 0 and 1, but since 0.5 corresponds to random guessing, no realistic classifier should be below 0.5 [11]. A ROC curve with corresponding AUC value can be seen in Figure 7.7. The remaining ROC curves can be seen in Appendix B. All AUC scores with corresponding p-values are stated in Tables 7.3 and 7.4. All combinations of diagnostic groups with the different methods are significant except cross-sectional FreeSurfer bam12 AD vs. MCI. Here, only an AUC of 0.52 is achieved, which almost corresponds to random guessing. To compare AUCs between methods (derived from the same subjects between diagnostic groups), DeLong test is used [10]. The test takes into account the correlated nature of the data. N-L Patch AUCs are compared to cross-sectional FreeSurfer and Longitudinal FreeSurfer. Significance between methods, indicated by *, ** and *** in Tables 7.3 and 7.4, are found for N-L patch and cross-sectional FreeSurfer bam12 for diagnostic groups AD vs. CN and AD vs. MCI and bam24 AD vs. MCI. No significant difference is found between N-L Patch and longitudinal FreeSurfer.

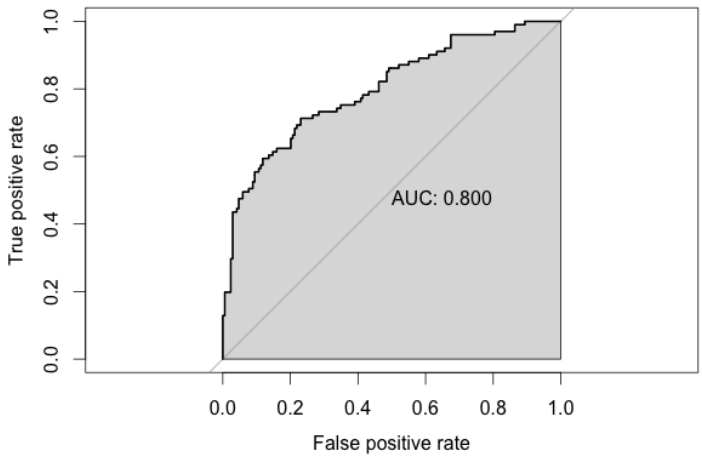


Figure 7.7: ROC curve. AD vs. CN for N-L Patch bam12. True positive rate vs. false positive rate as a function of the threshold of the measure that is evaluated.

Cohens' D:
Cohens' D is a measure of effect size which indicates the strength of a phenomenon. It is defined as the difference between two means, μ_1 and μ_2, divided by a measure based on the standard deviations, σ_1 and σ_2, for the data.

$$Cohens'D = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2+\sigma_2^2}{2}}}, \tag{7.3}$$

Cohens' D are correlated with another statistical measure typically used in drug testing, sample size. A low Cohens' D indicates the necessity of a larger sample size. Bootstrapping is used to compare Cohens' Ds between methods. To compute the p-value for bootstrapping, a two-tailed t-test for the null hypothesis of equal measures N_1-N_2=0 is carried out, where N_1 and N_2 are independent random measures. A probability distribution is computed for the difference between the Cohens' D for the two measures and computes the p-value as $p(N_1 > N_2) = 1 - \text{cdf}_{N_1-N_2}(0)$ and $p(N_2 > N_1) = 1 - p(N_1 > N_2)$, where cdf is the cumulative distribution function.

N-L Patch Cohen's Ds are compared to cross-sectional FreeSurfer and Longitudinal FreeSurfer. Significance between methods, indicated by ** and *** in Table 7.3 and by * in Table 7.4, are found for N-L patch and cross-sectional FreeSurfer bam12 and bam24 for diagnostic groups AD vs. CN and AD vs. MCI. No significant difference is found between N-L Patch and longitudinal FreeSurfer.

7.3.1 Linear regression

In the following model, it is assumed that the hippocampal percentual loss per year is constant. This means that the volume loss is an exponentially decaying function. To find the rate, a straight line is fitted by least squares estimates [23] to three points calculated as log(Follow-up volume/Baseline volume). At baseline, the follow-up volume is equal to the baseline volume, accordingly, the first point is always 0. The best fitted exponential line to the three target points is found, where best is defined by the least squares estimates. The fitted lines for the three methods can be seen in Figures 7.8, 7.9 and 7.10.

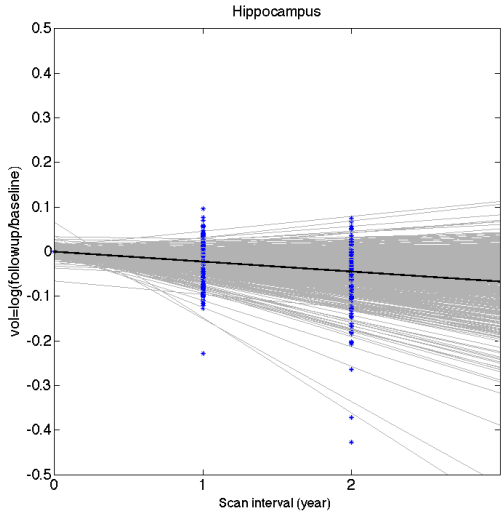


Figure 7.8: N-L Patch: Blue dots: Volume, (log(follow-up/baseline)), as a function of scan interval. Grey lines: Best fitted lines to three time points using least squares estimates. Black line: Mean.

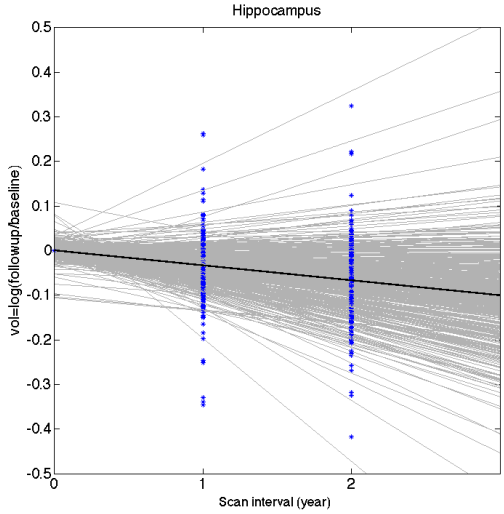


Figure 7.9: FS Cross: Blue dots: Volume, (log(follow-up/baseline)), as a function of scan interval. Grey lines: Best fitted lines to three time points using least squares estimates. Black line: Mean.

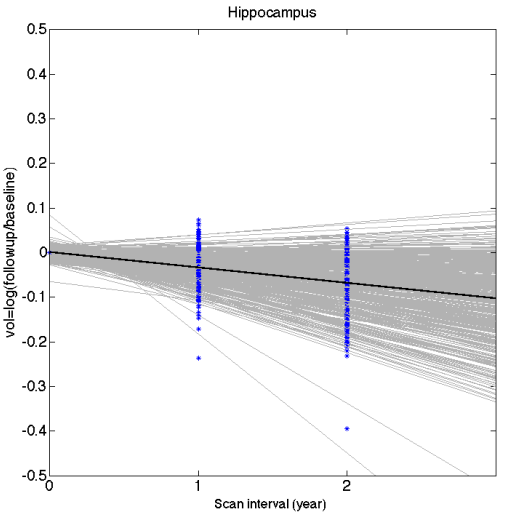


Figure 7.10: FS long: Blue dots: Volume, (log(follow-up/baseline)), as a function of scan interval. Grey lines: Best fitted lines to three time points using least squares estimates. Black line: Mean.

The rate per year for each subject is found by:
rate (%) = (exp(slope coefficient) - 1)*100. The mean rate for each diagnostic group with N-L Patch, cross-sectional FreeSurfer and longitudinal FreeSurfer can be seen in Table 7.5.

	AD	**MCI**	**CN**
	mean $\pm\sigma$	mean $\pm\sigma$	mean $\pm\sigma$
N-L Patch	-3.88 $\pm$2.63	-2.32 $\pm$2.45	-0.95 $\pm$1.32
FS Cross	-4.88 $\pm$3.58	-3.71 $\pm$3.31	-1.62 $\pm$2.59
FS Long	-5.36 $\pm$2.74	-3.74 $\pm$3.11	-1.61 $\pm$1.62

Table 7.5: Statistics based on rate/year for ADNI504. The rate is found by fitting a straight line to 3 time points calculated as log(follow-up volume / baseline volume) using least squares estimates.

From Figures 7.8, 7.9 and 7.10 it can be seen that the deviation looks larger for cross-sectional FreeSurfer than N-L Patch and longitudinal FreeSurfer. According to the model, the lines found by least squares estimates should intersect

(0,0). To test for bias, a two sided one sample t-test is used to test if the lines intersections with the y-axis for each method is equal to zero. The result of the test can be seen in Table 7.6.

One sample t-test			
	N-L Patch	FS Cross	FS Long
t-vale (p-value)	-0.79 (0.432)	0.46 (0.649)	2.19 (0.029)

Table 7.6: Two sided one sample t-test for each method based on the intersection with the y-axis of best fitted straight lines to three time points using least squares estimates. Volume is defined as log(Volume Follow-up/Volume Baseline).

From Table 7.6 it can be seen, that the null hypothesis, $\mu = 0$, can be rejected for FS long. This can be due to bias, uncertainties in measurements or a wrong model assumption - the hippocampal volume is not an exponentially decaying function.

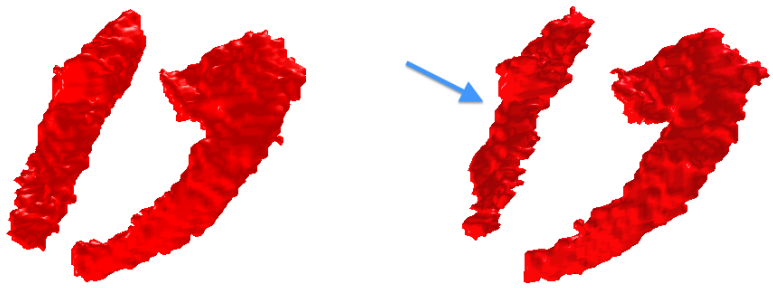
7.4 Discussion

As stated in Section 7.2, the initial segmentations are insufficient in some cases (total volume below 3500 voxels). It is not unusual that inter-subject registrations fail, since large intensity differences between subjects can lead to a scenario where the optimization can be caught at a local minimum as explained in Chapter 4. However, when the skull-stripping is changed, the segmentations are good in all cases. This could indicate the need to inspect all the skull-stripped images prior to segmentation or simply use another type of preprocessed images, e.g. bias field corrected images. The argumentation for using skull-stripped images in the first place is that the registration is not dominated by the high intensity skull, thereby resulting in a better hippocampal alignment. A precise hippocampal alignment is important to achieve good segmentation results since N-L Patch only search for similar patches within a $9\times9\times9$ voxels search volume. Therefore, changing the input images to bias field corrected images might not improve registration and thereby segmentation.

Segmentation with N-L Patch is in the same space after an affine inter-subject registration, since all atlases and subjects are registered to one atlas. However, when the inter-subject registration is changed to a rigid registration to the same

atlas, no scaling and shearing is done, which means the segmentations results are available in another space. Thus, the segmentations cannot directly be compared in these cases, and must be taken to a standard space. Furthermore, doing a comparison with the other methods involves taking the segmentations to subject FreeSurfer space and doing an interpolation. This affects Dice scores with the manual labels, which changes from 0.868 ± 0.019 to 0.855 ± 0.021 for Atlas40, Chapter 6. This could be avoided if a computational heavier alternative was used where atlases were registered to the test subject, as done in BrainFuseLab and Section 6.1.1, and might accordingly be considered in future work.

From Tables 7.3 and 7.4 it can be seen that the standard deviations for N-L Patch is increased for bam24 compared to bam12. This can be due to more extreme outliers which can be seen from the histograms in Appendix B. These outliers are not corrected, since the total hippocampal volume is not below 3500 voxels. Figure 7.11 illustrates a segmentation where the left hippocampus of an AD subject, indicated by blue arrow, is not segmented properly, which leads to a atrophy score from baseline to m24 of -31 %. The left hippocampus volume decreases from 3867 to 1906 mm^3, whereas the right decreases from 4562 to 3909 mm^3. Since the left and right hippocampus should have approximately the same volume, the fraction between the volumes could have been used as an indicator to decide if the MRI should be inspected for e.g. an imprecise skull-stripping. Figure 7.12 illustrates the transformed MRI, green, superimposed on the test atlas, greyscale. In this case, it does not look like a faulty registration. More likely, it is not possible to use affine registration to match the brain of this subject, because the pathological brain changes are very large, e.g. enlarged ventricles.



(a) Baseline (b) m24

Figure 7.11: N-L Patch segmentation of 136_S_0426 at baseline and m24 where total atrophy is estimated to -31%. Arrow indicates the left hippocampus, which during two years has decreased 49 % in volume.

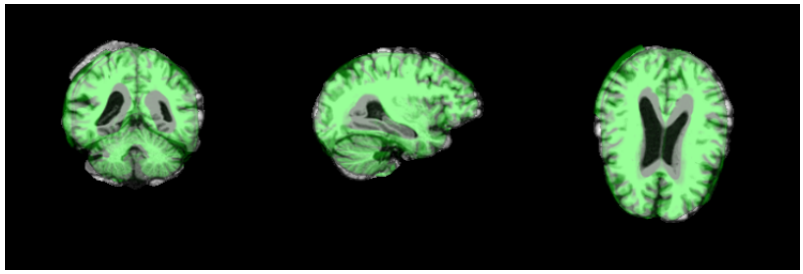


Figure 7.12: 136_S_0426 registration m24. Warped subject MRI (green) superimposed on test subject (greyscale).

N-L Patch yielded significantly better group separation than cross-sectional FreeSurfer in separating AD from CN and AD from MCI for bam12 and bam24, based on bootstrapping. Furthermore, N-L Patch also yielded significantly better group separation than cross-sectional FreeSurfer in separating AD from CN and AD from MCI for bam12 and AD from MCI for bam24, based on DeLong test. Longitudinal FreeSurfer exploiting baseline and follow-up simultaneously was tested and showed no diagnostic improvement over N-L Patch when doing bootstrapping and Delong test between methods.

In N-L Patch, two fundamental aspects have been changed compared to cross-sectional FreeSurfer: the method and the atlas. It is difficult to decide, if the better results obtained with N-L Patch are due to the one or the other. To make

a fair comparison, the methods should preferably use the same atlas. However, since there is no access to the atlases used to build the probabilistic FreeSurfer atlas, this has not been possible.
The new Harmonized Hippocampal Protocol (HHP) label definition is made by many experts after evaluating a variety of segmentation protocols and then agreeing on an equivocal definition. This suggests that the manual HHP segmentations are more standardized and thereby better than the manual segmentations used in FreeSurfer. Furthermore, HHP labels are considered the new hippocampal golden standard. The cross-subject averaging done to make the FreeSurfer atlas might have removed useful information. On the other hand, the FreeSurfer atlas has been used for a long time, and the atlases used to make this probabilistic atlas are used in other automated multi-atlas methods as well e.g. [28], which illustrates their robustness in automated segmentation.

FreeSurfer and N-L Patch methods are fundamentally different. FreeSurfer uses Markov Random Fields in a Baysian framework to do segmentation. Information from intensity, prior probabilities at a voxel and labels of neighboring voxels, is included in the model.
In its original form, the N-L patch method results in segmentations with speckle patterns, as illustrated in Section 6.1. These areas are wrongly classified as part of hippocampus because the method does not take labels of neighboring classified voxels into account. However, the speckles can easily be removed automatically as done in this thesis. The segmentations obtained using FreeSurfer are often rough and has branches of misclassified voxels, Figure 1.7. Since these branches are not represented in the FreeSurfer atlas, they arise because of the method. The branches are interconnected with the hippocampus border and can only be removed by morphological operations, which effects the entire surface volume. Since the surface volume accounts for approximately 10 % of the total volume, this is not an optimal solution when estimating atrophy.

N-L Patch only uses affine registrations or the subclass, rigid registrations, and is therefore considerably faster than methods using non-linear registration. No general rule exist, but methods that uses a lot computational power are often more precise than methods that use less. The question is if N-L Patch only performs well because a good atlas from the HHP is available? In [8], N-L Patch is used for ventricle segmentation with another atlas and gives good results. Segmentation of 80 elderly subjects with mild to moderate AD gives a mean Dice score = 0.959. This indicates the possibility to extend the method to other structures. However, the computational cost increases with structure size. To make N-L Patch even faster in order to get segmentation results immediately after the scan, the images could be cropped and the method could be implemented on a GPU.

Chapter 8

Conclusion

The ambition behind this thesis was to improve automated segmentation of hippocampus from T1-weighted Magnetic Resonance Imaging (MRI) compared to the current method (FreeSurfer) used at the company *Biomediq A/S*. Biomediq A/S strives at eliminating the use of FreeSurfer in their alzheimer's diagnostic pipeline.

In an initial literature study, multi-atlas segmentation methods were found to be among the top performing automated hippocampal segmentation methods. These methods rely on manual annotations called atlases. Two fundamentally different multi-atlas methods were chosen, to analyze if the best performance was achieved with 1) the relatively faster *N-L Patch* using affine registrations to align MRIs and atlases or with 2) a computationally heavier method, *BrainFuseLab*, using non-rigid registrations to align atlases and MRIs. In N-L Patch a label is obtained for every voxel by using similar image patches from coarsely aligned atlases, whereas BrainFuseLab gives atlases with local similarity to the test subject high weight when a voxel is labeled. For both methods, manual annotations from a new Harmonized Hippocampal Protocol (HHP) were used as atlases. These manual annotations include both subjects with alzheimer's disease (AD), mild cognitive impairment (MCI) and cognitively normal (CN), and are furthermore considered the hippocampal golden standard.

Before segmentation the MRI data were preprocessed in several steps for both

methods. This involved removal of spatial intensity inhomogeneities, skull-stripping and transformation of atlases and MRIs to one segmentation space.

Method parameters were optimized in a leave-one-out cross-validation using two different HHP atlas sets. A paired t-test between Dice scores of N-L Patch and BrainFuseLab, showed significance ($p<0.001$), accordingly N-L Patch yielded significantly higher Dice scores than BrainFuseLab. This illustrates, that heavy computational methods not necessarily give better results than fast methods. Based on Dice scores, computation time and visual inspection, N-L Patch was chosen to be the optimal method. Furthermore, N-L Patch resulted in better segmentation consensuses with the new hippocampal label standard than the state-of-the-art method, cross-sectional FreeSurfer.

N-L Patch was used to segment a standardized ADNI dataset containing 1.5T MRIs from 504 subjects (169 CN, 234 MCI, 101 AD) at baseline, month 12 and month 24. In cases, where the registration was considered to fail, either the skull-stripping was redone or the registration was changed to a rigid registration as it was considered that the scaling caused the fail. After these adjustments, no insufficient segmentations were achieved with hippocampal volume below 3500 mm^3.

Hippocampal atrophy rate calculated as percentage volume change from baseline to follow-up was estimated. Based on a statistical analysis, the diagnostic group separation capabilities of N-L Patch were compared to two state-of-the-art methods, 1) cross-sectional FreeSurfer and 2) longitudinal FreeSurfer. Including the HHP labels in N-L Patch yielded significantly better group separation than cross-sectional FreeSurfer in separating AD from CN and AD from MCI. Also longitudinal FreeSurfer exploiting baseline and follow-up simultaneously showed no diagnostic improvement over N-L Patch. This illustrates the longitudinal robustness of segmentations when annotations from the new hippocampal label standard are included in automated segmentation methods.

Two fundamentally different aspects were changed in N-L Patch compared to the FreeSurfer methods: the method and the atlas. The definition of the HHP labels have been agreed upon by many experts, which indicates that the atlases used in N-L Patch are better than the probabilistic atlas used in FreeSurfer, where the averaging of atlases might have removed usefull information. Based on the visual segmentations, both N-L Patch and FreeSurfer have areas of misclassified voxels. However, in N-L Patch these can easily be removed since they are not part of the hippocampal border which is not the case in FreeSurfer. This indicates, that the method used in N-L Patch results in better automated segmentations than FreeSurfer.

Based on the results of this thesis, Biomediq A/S now has achieved a fast

hippocampal segmentation method independent of the segmentation part of FreeSurfer. Additionally, Biomediq A/S now has a method with full access to the source code. However, the use of FreeSurfer is not eliminated as the MRIs still need to be preprocessed. Developing a preprocessing pipeline might be considered in future work, if the use of FreeSurfer should be eliminated completely. If atlases are available for other subcortical structures, the N-L Patch method can easily be extended to segment such structures as well. A GPU implementation will decrease computation time further, with high clinical relevance, since segmentations should preferably be available in the clinic immediately after scanning.

Overall, hippocampal N-L Patch segmentation with HHP labels showed convincing results, though, there is still room for improvements and new features in the segmentation.

8.1 Future Work

Below is a list with possible elements for improving hippocampal N-L Patch segmentations and steps in getting a better subcortical segmentation pipeline.

1. Include a similarity measure based on texture analysis of the test and training subjects in the non-local means label fusion, Equation 5.2.
2. If many similar patches are found from the same training subject, it indicates that this subject globally looks like the test subject. This could be incorporated in the label fusion, so patches from that specific training subject weights more in the final fusion of labels, Equation 5.2.
3. Optimize N-L patch code by cropping images around the structure of interest and make a GPU implementation.
4. Extent the method to other structures. This involves finding good atlas sets.

Appendix A

Atlas Demographics

This appendix includes the demographics of the manual segmentations from the Harmonized Hippocampal Protocol used as atlas sets in this thesis. One atlas set contains 15 atlases, Table A.1, the other contains 40 atlases, Table A.2. MMSE score is explained in Chapter 3.

PatientID	SeriesID	Diagnosis	Age	Gender	MMSE	Hand
002_S_0816	29.612	AD	71,4630	1	26	1
003_S_0931	20.050	CN	86,2247	2	28	1
003_S_1059	22.301	AD	84,5616	2	25	1
003_S_1257	27.340	AD	85,1342	1	20	1
009_S_0842	18.870	CN	73,7151	1	28	1
009_S_0862	19.358	CN	73,4685	2	30	1
009_S_1030	21.823	MCI	67,5507	1	28	1
009_S_1334	50.567	AD	64,8767	1	22	1
011_S_0002	9.107	CN	74,3863	1	28	2
013_S_0592	18.419	AD	78,1370	1	23	1
013_S_1276	27.641	CN	71,9644	2	30	1
016_S_1092	23.826	MCI	74,4685	1	26	1
016_S_1263	27.304	AD	64,8822	2	26	1
100_S_1062	66.023	AD	84,4712	1	28	1
100_S_1286	64.890	CN	77,7205	2	28	2

Figure A.1: Atlas15 demographics. Gender: Male = 1, female = 2. Hand: Right = 1, left = 2.

PatientID	SeriesID	Diagnosis	Age	Gender	MMSE	Hand
002_S_0295	13.408	CN	84,8904	1	28	1
002_S_0413	13.893	CN	76,3836	2	29	1
002_S_0782	17.835	MCI	81,7205	1	29	1
002_S_0816	29.612	AD	71,4630	1	26	1
002_S_0938	19.852	AD	82,3068	2	23	1
003_S_0907	19.728	CN	88,7260	2	30	1
003_S_0908	32.516	MCI	62,9616	2	29	1
003_S_0931	20.050	CN	86,2247	2	28	1
003_S_1057	23.345	MCI	61,3753	2	26	1
003_S_1059	22.301	AD	84,5616	2	25	1
003_S_1122	23.542	MCI	76,8055	2	28	1
003_S_1257	27.340	AD	85,1342	1	20	1
005_S_1341	27.673	AD	71,7151	2	24	1
007_S_0101	10.679	MCI	73,6356	1	27	1
007_S_0128	10.936	MCI	64,1288	2	29	1
009_S_0842	18.870	CN	73,7151	1	28	1
009_S_0862	19.358	CN	73,4685	2	30	1
009_S_1030	21.823	MCI	67,5507	1	28	1
009_S_1334	50.567	AD	64,8767	1	22	1
010_S_0067	10.344	CN	74,5562	1	27	1
011_S_0002	9.107	CN	74,3863	1	28	2
011_S_0010	8.800	AD	73,9699	2	24	1
011_S_0016	9.253	CN	65,4630	1	28	1
011_S_0183	12.000	AD	72,5452	2	21	1
011_S_0856	19.031	AD	60,3781	1	27	2
013_S_0592	18.419	AD	74,0000	1	28	1
013_S_1276	27.641	CN	71,9644	2	30	1
016_S_1092	23.826	MCI	74,4685	1	26	1
016_S_1263	27.304	AD	64,8822	2	26	1
098_S_0149	11.021	AD	87,8137	1	20	1
098_S_0172	11.812	CN	70,6384	2	29	1
100_S_0995	66.038	AD	81,0548	2	26	1
100_S_1062	66.023	AD	84,4712	1	28	1
100_S_1286	64.890	CN	77,7205	2	28	2
123_S_0050	10.053	MCI	77,7233	1	26	2
123_S_0091	15.898	AD	62,9616	1	25	1
123_S_0094	15.867	AD	71,3014	2	20	1
123_S_1300	27.689	MCI	73,5452	2	28	2
127_S_0259	12.137	CN	70,6301	1	30	1
127_S_0754	18.515	AD	67,7123	2	23	1

Figure A.2: Atlas40 demographics. Gender: Male = 1, female = 2. Hand: Right = 1, left = 2.

Appendix B

Statistical Analysis

The results in this appendix are obtained from a statistical analysis of the final segmentations from *ADNI504* with Non-Local Patch-based segmentation (N-L Patch), cross-sectional FreeSurfer (FS cross) and longitudinal FreeSurfer (FS long) explained in Chapter 7.

B.1 Volume results

Tables B.1 and B.2 states the mean $\pm\sigma$ for N-L Patch, cross-sectional FreeSurfer and longitudinal FreeSurfer for the diagnostic groups at m12 and m24, respectively. The corresponding baseline table is Table 7.2. Since one patient is not diagnosed at m24, the results are obtained using the baseline diagnostics.

	Mean volume $(mm^3) \pm \sigma$		
	CN(n=169)	MCI(n=234)	AD(n=101)
N-L Patch	7792 ± 889	6996 ± 1096	6247 ± 1182
FS Cross	7107 ± 1018	6075 ± 1110	5330 ± 1197
FS Long	7055 ± 1052	5921 ± 1173	5105 ± 1276

Table B.1: ADNI504 m12: Mean hippocampal volume for diagnostic groups with different methods.

	Mean volume $(mm^3) \pm \sigma$		
	CN(n=169)	MCI(n=234)	AD(n=101)
N-L Patch	7716± 908	6850 ± 1149	6034 ± 1223
FS Cross	6985 ± 1014	5869 ± 1161	5049 ± 1195
FS Long	6918 ± 1059	5702± 1220	4817 ± 1253

Table B.2: ADNI504 m24: Mean hippocampal volume for diagnostic groups with different methods.

Scatter plots of N-L Patch volume against cross-sectional FreeSurfer or longitudinal FreeSurfer at m12 and m24 can be seen in Figures B.1 and B.2, respectively. The corresponding figures for baseline can be seen in Figures 7.4 and 7.5. Since one patient is not diagnosed at m24, the results are obtained using the baseline diagnostics.

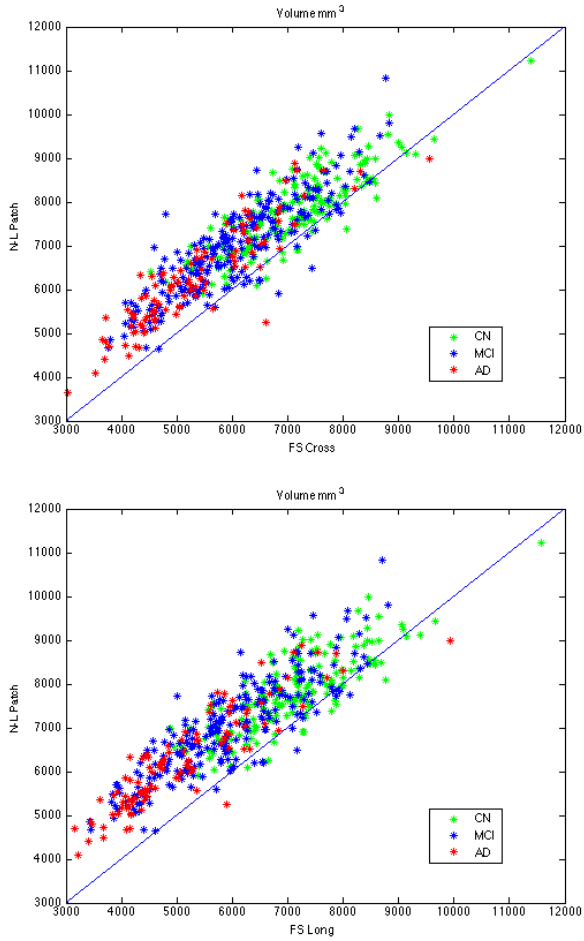


Figure B.1: ADNI504 m12: Scatter plots. Top: Hippocampal volume with N-L Patch vs. FS Cross. Bottom: Hippocampal volume with N-L Patch vs. FS Long.

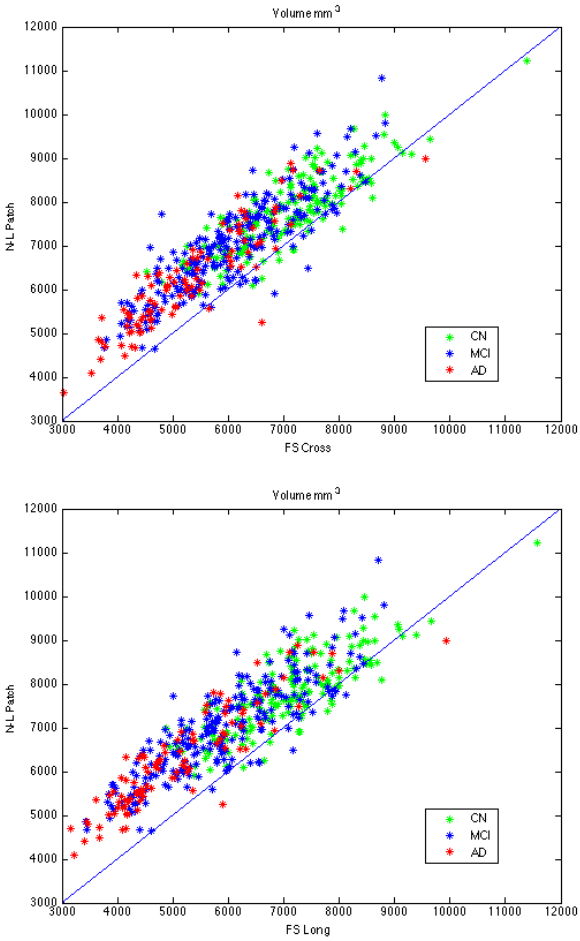
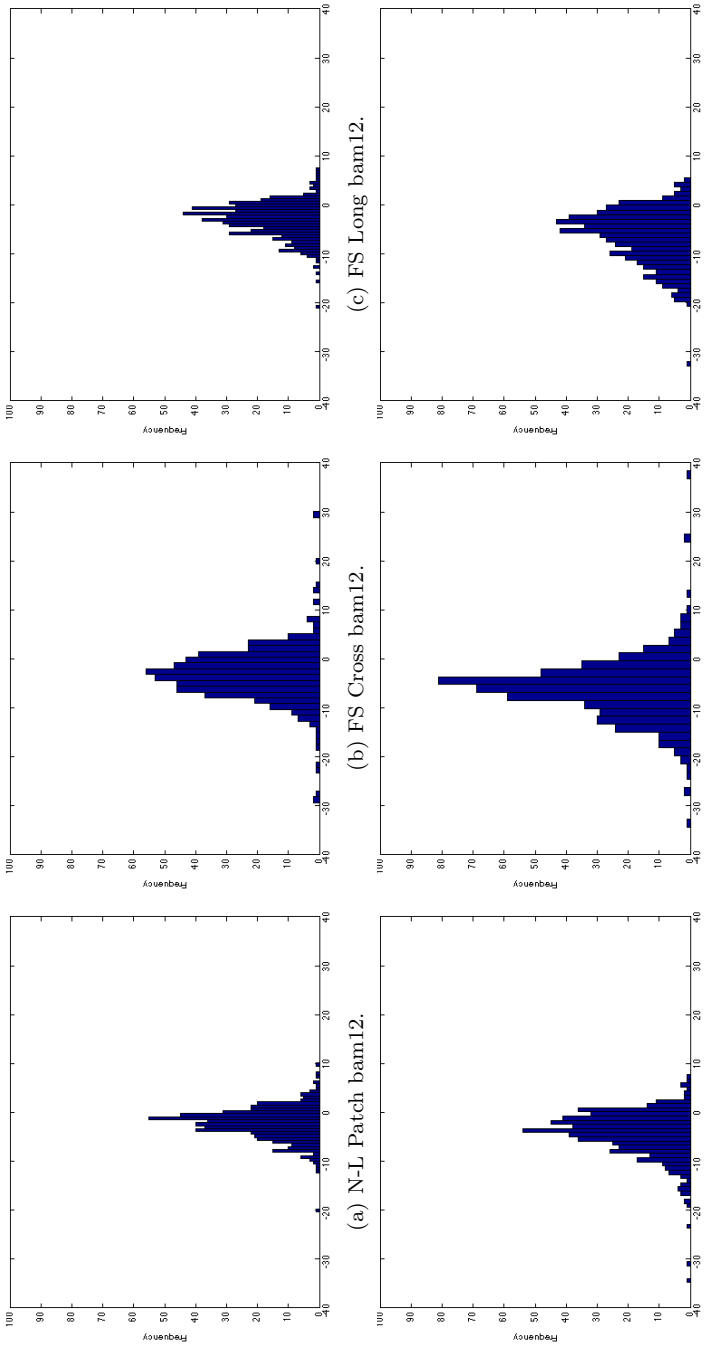


Figure B.2: ADNI504 m24: Scatter plots. Top: Hippocampal volume with N-L Patch vs. FS Cross. Bottom: Hippocampal volume with N-L Patch vs. FS Long.

B.2 Atrophy histograms



(a) N-L Patch bam12. (b) FS Cross bam12. (c) FS Long bam12.

(d) N-L Patch bam24. (e) FS Cross bam24. (f) FS Long bam24.

Figure B.3: Frequency vs. percentage difference in hippocampal volume (atrophy) between baseline and followup (bam12 or bam24) with different methods.

B.3 Bartlett's Test

To test for variance homogeneity between groups, Bartlett's test, is performed on atrophy rates for the three diagnostic groups, AD, MCI and CN for each method. The results of Bartlett's test are shown in Tables B.3 and B.4. The low p-values reveal variance inhomogeneity between the clinical diagnostic groups in N-L Patch method and longitudinal FreeSurfer for both bam12 and bam24 and cross-sectional FreeSurfer bam24. Variance homogeneity can not be rejected for cross-sectional FreeSurfer bam12.

Segmentation Method	p-value
N-L Patch	<0.001
FS Cross	0.9469
FS Long	<0.001

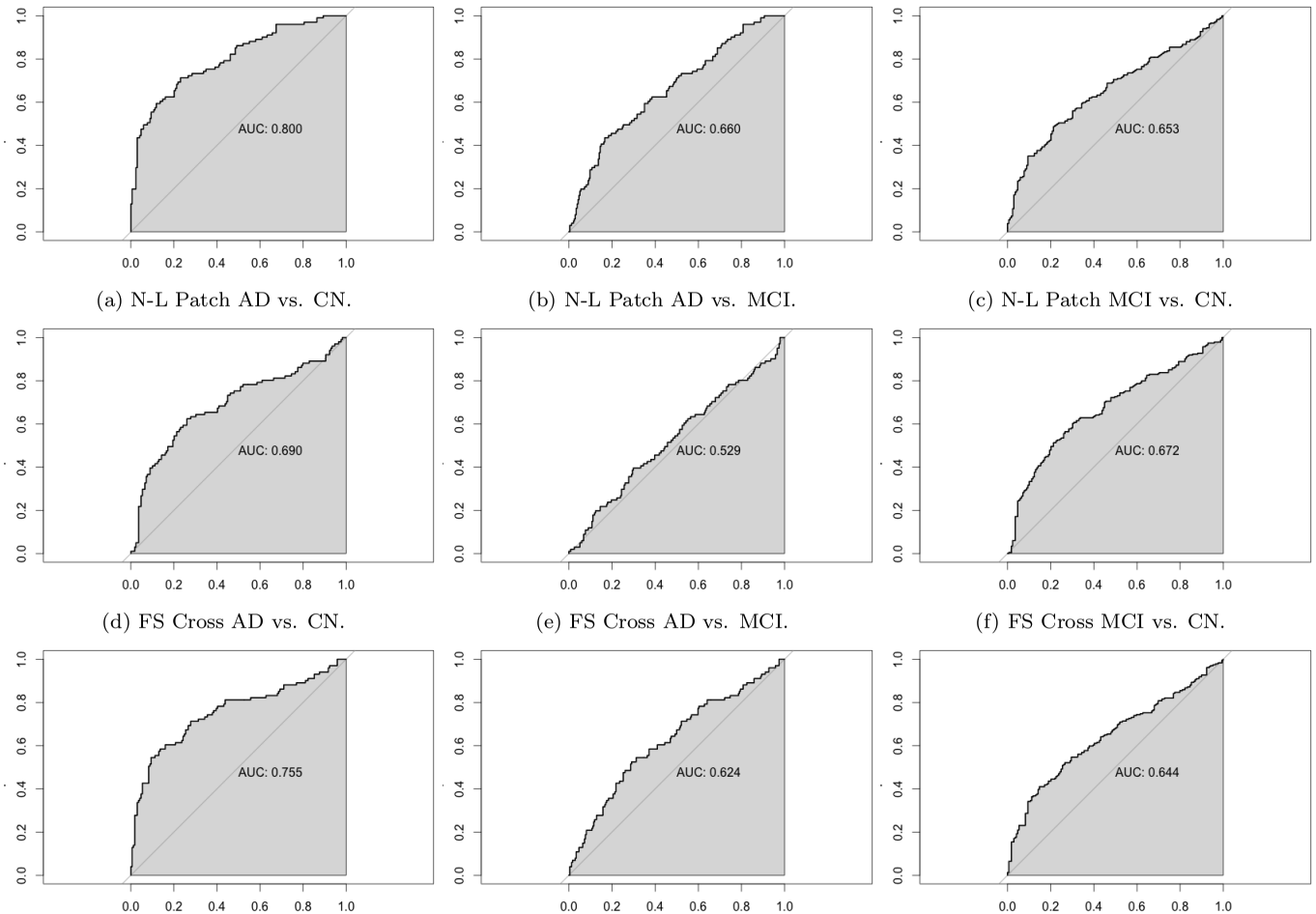
Table B.3: Bartlett's tests of inhomogeneity of variances bam12.

Segmentation Method	p-value
N-L Patch	<0.001
FS Cross	0.007
FS Long	<0.001

Table B.4: Bartlett's tests of inhomogeneity of variances bam24.

B.4 ROC curves

ROC curves for bam12 and bam24 with corresponding AUC values can be seen in Figures B.4 and B.5.

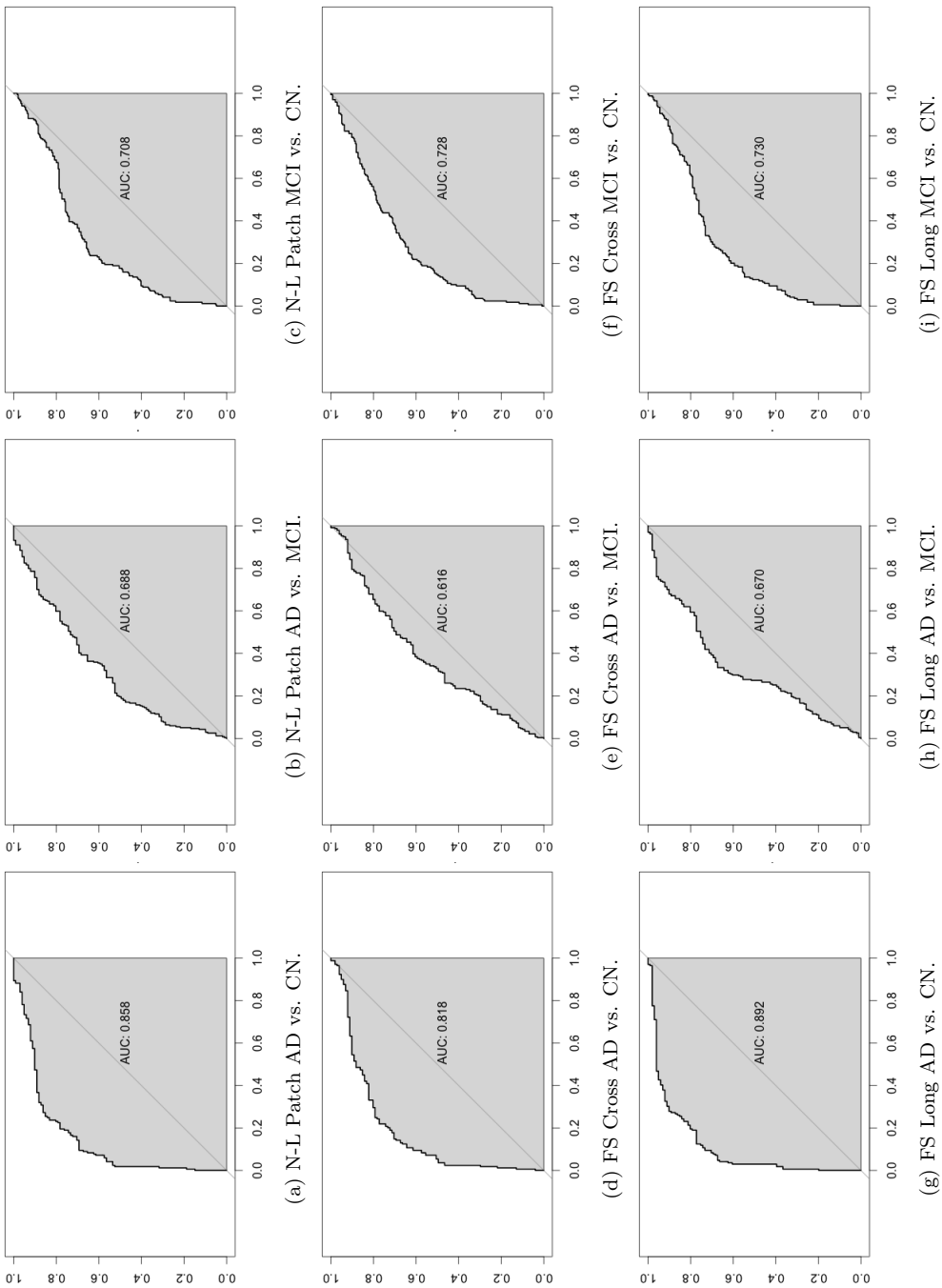


(a) N-L Patch AD vs. CN. (b) N-L Patch AD vs. MCI. (c) N-L Patch MCI vs. CN.

(d) FS Cross AD vs. CN. (e) FS Cross AD vs. MCI. (f) FS Cross MCI vs. CN.

(g) FS Long AD vs. CN. (h) FS Long AD vs. MCI. (i) FS Long MCI vs. CN.

Figure B.4: ROC curves bam12.



(a) N-L Patch AD vs. CN. (b) N-L Patch AD vs. MCI. (c) N-L Patch MCI vs. CN.

(d) FS Cross AD vs. CN. (e) FS Cross AD vs. MCI. (f) FS Cross MCI vs. CN.

(g) FS Long AD vs. CN. (h) FS Long AD vs. MCI. (i) FS Long MCI vs. CN.

Figure B.5: ROC curves bam24.

Appendix C

Data CD

The enclosed CD contains the following:

- Volume results of ADNI504 at 3 time points with N-L Patch, FS cross and FS long.
- Atrophy scores of ADNI504 (bam12, bam24) with N-L Patch, FS cross and FS long.
- N-L Patch source code.
- R-code and m-code used for statistics.

Bibliography

[1] P. Aljabar, R.A. Heckemann, A. Hammers, J.V. Hajnal, and D. Reuckert. Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. *Neuroimage*, 46:726–738, 2009.

[2] John Ashburner and Karl. J. Friston. Rigid body registration. in: statistical parametric mapping: The analysis of functional brain images. *Academic Press*, pages 49–62, 2007.

[3] Dubois B, Feldman HH, Jacova C, DeKosky ST, , Delacourte A Barberger-Gatteau P, Cummings J, Galalsko D, Gauthier S, Jicha G, Meguro K, O'Brian J, Pasquier F, Robert P, Rossor M, Salloway S, Stern Y, Visser Pj, and Scheltens P. Research criteria for the diagnosis of alzheimer's disease: revising the nincds-adrda criteria. *Lancet Neuron*, 43(6):734–746, 2007.

[4] Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A, Killiany R, Kennedy D, Klaveness S, Montillo A, Makris N, Rosen B, and Dale AM. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–355, 2002.

[5] Kolawole Oluwole Babalola, Brian Patenaude, Paul Aljabar, Julia Schnabel, David Kennedy, William Crum, Stephen Smith, Tim Cootes, Mark Jenkinson, and Daniel Rueckert. An evaluation of four automatic methods of segmenting the subcortical structures in the brain. *Neuroimage*, 47:1435–1447, 2009.

[6] Marie Chupin, Emilie Gérardin, Rémi Cuignet, Claire Boutet, Louis Lemieux, Stéphane Lehéicy, Habib Benali, Line Garnero, Olivier Colliot,

and the Alzheimers Disease Neuroimaging Initiative. Fully automatic hippocampus segmentation and classification in alzheimer's disease and mild cognitive impairment applied on data from adni. *Hippocampus*, 19(6):579–787, 2009.

[7] Pierrick Coupé, Simon F. Eskildsen, José V. Manjón, Vladimir Fonov, D. Louis Collins, and the Alzheimer's Disease Neuroimaging Initiative. Simultaneous segmentation and grading of anatomical structures for patient's classification: Application to alzheimer's disease. *NeuroImage*, 59(4):3736–47, 2012.

[8] Pierrick Coupé, José V. Manjón, Vladimir Fonov, Jens Pruessner, Montserrat Robles, and D. Louis Collins. Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. *NeuroImage*, 54:940–954, 2011.

[9] Anders M. Dale, Bruce Fischl, and Martin I Sereno. Cortical surface-based analysis. i. segmentation and surface reconstruction. *Neuroimage*, 9:179–194, 1999.

[10] Elizabeth R. Delong, David M. Delong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, 1988.

[11] Tom Fawcett. http://home.comcast.net/ tom.fawcett/public_html/papers/roc101.pdf. 2004.

[12] Frisoni GB and Jack CR. Harmonization of magnetic resonance-based manual hippocampal segmentation: A mandatory step for wide clinical use. *Alzheimer's and Dementia*, 7:171–174, 2011.

[13] A. Ghanei, H. Soltanian-Zadah, and J.P. Windham. Segmentation of the hippocampus from mri using deformable contours. *Comput. Med. Imaging. Graph*, 22(3):203–216, 1998.

[14] Rolf A. Heckemann, Joseph V. Hajnal, Poul Aljabar, Daniel Rueckert, and Alexander Hammers. Automatic anatomical brain mri segmentation combining label propagation and decision fusion. *NeuroImage*, 33:115–126, 2006.

[15] Rolf A. Heckemann, Shiva Keihaninejad, Paul Aljabar, Daniel Rueckert, Joseph V. Hajnal, and Alexander Hammers. Improving intersubject image registration using tissue-class information benefits robustness and accuracy of multi-atlas based anatomical segmentation. *NeuroImage*, 51:221–227, 2010.

[16] http://itk.org.

[17] http://people.csail.mit.edu/msabuncu/.

[18] http://picsl.upenn.edu/Project/MALF.

[19] http://surfer.nmr.mgh.harvard.edu/.

[20] http://www.alz.org/downloads/facts_figures_2012.pdf.

[21] http://www.fil.ion.ucl.ac.uk/spm/software/.

[22] http://www.hippocampal protocol.net.

[23] Richard Johnson, John Freund, and Irwin Miller. Miller and freund's probability and statistics for engineers. eight edition. *Pearson*, page 304, 2013.

[24] Jack CR Jr, Knopman DS, Jagust WJ, Shaw LM, AIsen PS, Weiner MW, Petersen RC, and Trojanowski JQ. Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade. *Lancet Neurol*, 9(1):119–28, 2010.

[25] Kelvin K. Leung, Josephine Barnes, Gerard R. Ridgway, Jonathan W. Bartlett, Matthew J. Clarkson, Kate Macdonald, Norbert Schuff, Nick C. Fox, and Sebastian Ourselin. Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and alzheimer's disease. *Neuroimage*, 51(4):1345–1359, 2010.

[26] West MJ, Coleman PD, Flood DG, and Troncoso JC. Differences in the pattern of hippocampal neuronal loss in normal ageing and alzheimer's disease. *Lancet*, 344(8925):769–772, 1994.

[27] Sean M. Nestor, Eric Gibson, Fu-Qiang Gao, Alex Kiss, and Sandra E. Black. A direct morphometric comparison of five labeling protocols for multi-atlas driven automatic segmentation of hippocampus in alzheimer's disease. *NeuroImage*, 66:50–70, 2013.

[28] Mert R. Sabuncu, B. T. Thomas, Koen Van Leemput, Bruce Fischl, and Polina Golland. A generative model for image segmentation based on label fusion. *IEEE Transactions on Medical Imaging*, 29(10):1714–1729, 2010.

[29] F. Ségonne, A.M. Dale, E. Busa an M. Glessner, D. Salat, H.K. Hahn, and B. Fischl. A hybrid approach to the skull stripping problem in mri. *Neuroimage*, 22:1060–1075, 2004.

[30] John G. Sled, Alex P. Zijdenbos, and Allen C. Evans. A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE Transactions on Medical Imaging*, 17:87–97, February 1998.

[31] Fedde van der Lijn, Tom den Heijer, Monique M.B. Breteler, and Wiro J. Niessen. Hippocampus segmentation in mr images using atlas registration, voxel classification, and graph cuts. *NeuroImage*, 43:708–720, 2008.

[32] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholos Ayache. Symmetric log-domain diffeomorphic registration: A demons-based approach. *Springer. Lecture Notes Computer Science*, 5241:754–761, 2008.

[33] Robin Wolz, Paul Aljabar, Joseph V. Hajnal, Alexander Hammers, Daniel Rueckert, and the Alzheimer's Disease Neuroimaging Initiative. Leap: Learning embeddings for atlas propagation. *Neuroimage*, 49(2):1316–1325, 2010.

[34] B.T. Wyman, D. J. Harvey, K. Crawford, M. A. Bernstein, O. Carmichael, P. E. Cole, P. K. Crane, C. DeCarli, N. C. Fox, J. L. Gunter, D. Hill, R. J. Killiany, C. Pachai, A. J. Schwarz, N. Schuff, M. L. Senjem, J. Suhy, P. M. Thompson, M.l Weiner, Jr C. R. Jack, and for the Alzheimer's Disease Neuroimaging Initiative. Standardization of analysis sets for reporting results from adni mri data. *Alzheimer's and Dementia*, 9:332–337, 2013.